Li WEIGANG, Mayara Chew MARINHO, Denise Leyi LI, Vitor Vasconcelos DE OLIVEIRA, 2024. Six-Writings multimodal processing with pictophonetic coding to enhance Chinese language models. *Frontiers of Information Technology & Electronic Engineering*, 25(1):84-105. https://doi.org/10.1631/FITEE.2300384

Six-Writings multimodal processing with pictophonetic coding to enhance Chinese language models

Key words: Chinese language model; Chinese natural language processing (CNLP); Generative language model; Multimodal processing; Six-Writings

Corresponding author: Li WEIGANG
E-mail: weigang@unb.br
ORCID: https://orcid.org/0000-0003-1826-1850

Abstract

- We propose a framework called Six-Writings (Xu S, 1997) multimodal processing (SWMP) to enable direct integration of Chinese NLP (CNLP) with morphological and semantic elements. The first part of SWMP, known as Six-Writings pictophonetic coding (SWPC), is introduced with a suitable level of granularity for radicals and components, enabling effective representation of Chinese characters and words.

- We conduct several experimental scenarios. The results demonstrate that SWMP/SWPC methods effectively capture the distinctive features of Chinese and offer a promising mechanism to enhance CNLP with better efficiency.

Motivation

- The field of computer information science and technology, including most AI, ML, and LLMs, has been developed based on English and corresponding coding. Chinese language is essentially translated into a form that modern computers can understand.
- UNICODE serves as the foundation for Chinese informatization but is insufficient for effectively calculating the similarity between Chinese characters. To achieve accurate similarity results, CNLP needs specialized coding approaches.
- The lack of a standardized Chinese character coding suitable for LLMs has resulted in the absence of a unified presentation for prompting questions. Despite most research findings indicating high similarity, CNLP's actual results are less than ideal.

Unicode 十六进制码点范围	UTF-8 二进制
0000 0000 – 0000 007F	0xxxxxxx
0000 0080 – 0000 07FF	110xxxxx 10xxxxxx
0000 0800 – 0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx
0001 0000 – 0010 FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

I. Framework of SWMP

- The Six-Writings of Chinese characters is a multimodal way of understanding Chinese characters (Xu S, 1997), which was summarized by the ancients. It takes into account how human beings understand Chinese characters through their five senses.
- We propose a framework called Six-Writings multimodal processing (SWMP) to enable direct integration of Chinese NLP (CNLP) with morphological and semantic elements.
- This framework for Chinese language models consists of six parts: (1) pictophonetic, (2) pinyin, (3) property, (4) image, (5) audio/ video, and (6) understanding (word embedding).

Pictophonetic Pinyin

I. Framework of SWMP

- This comprehensive multimodal processing framework allows for a detailed representation of Chinese characters and words in the Six-Writings style, facilitating various language processing tasks and enabling a deeper understanding of their structure and attributes.
- In practical applications, different parts of the SWMP framework can be combined based on specific task requirements.
- The second part creates the Six-Writings pinyin code, while the first and third parts can be combined to form the Six-Writings semantic code. Furthermore, integrating the first, second, and fourth parts results in the comprehensive text/pronunciation/image processing of Six-Writings.

II. SWPC approach - pictophonetic property

- Chinese characters are logograms with a hierarchical structure consisting of strokes and components. Strokes are the basic units of character formation, and components are formed by combining strokes. It is estimated that over 90% of Chinese characters belong to the phono-semantic (形声) category (Zhang B, 2008).
- Phono-semantic characters are typically composed of a semantic component, namely the radical (形旁) and a phonetic component (声旁). The semantic component represents the overall meaning or category of the character and is commonly referred to as the radical component (Yeromiyan, 2023).
- The phonetic component, on the other hand, serves as a means to differentiate characters with similar meanings or pronunciations.

蹬瞪僜

澄噔嶝

Character by left and right components

- Most Chinese characters are formed by combining the left and right components.
- For example, character 横 (horizontal) has the WB letter code SAMW and the WB numeric code 14152534. According to the WB coding rule, the key position of the left radical 木 (wood) on the QWERTY-based keyboard is S, and its position code is 14. The code of the phonetic component 黄 (yellow) is 152534.
- The normalized and augmented WB number is 17193957, with a radical code of 17 and a phonetic code of 193957, totaling 8 digits.
- Another character 酮 (ketones) has the WB code SGMK, the WB numeric code 14112523, the normalized and augmented code 17113935, the radical code 17, and the component code 113935. The radical codes of 横 (horizontal) and 酮 (ketones) are the same.

http://life.chacuo.net/convertxuezi

Character by left and right components

- To avoid code duplication and enhance the digital representation of the pictophonetic features, we combine the advantages of WB and FC and modify the radical 木 (wood) of character 横 (horizontal) to 1749, while keeping the phonetic 黄 (yellow) unchanged at 193957. The combination of the normalized and augmented WB numerical code and FC numbers is called SWPC. In this case, character 横 (horizontal) is coded as 1749193957, with a total of 10 digits.
- Similarly, the FC number of character 酮 (ketones) is 17620, with the radical 酉 (unitary) having a code of 16 and the phonetic component 同 (same) having a code of 720. The SWPC is 1716113935, effectively distinguishing the two radicals 木 (wood) and 酉 (unitary). This demonstrates the advantages of using SWPC to represent Chinese characters, especially when calculating the similarity between Chinese characters (or words) and conducting multimodal text/image processing.

III. SWPC for text/image processing

- SWPC provides convenience for multimodal processing of Chinese characters using image and text data. Drawing from the coding process of WB, tens of thousands of Chinese characters can be categorized into three main situations based on their character formation rules:
  - The character itself serves as a root character.
  - The character’s root is a radical (semantic) of another Chinese character, represented by a radical code (a part of SWPC).
  - The character’s root is a component (phonetic) of another Chinese character, represented by a phonetic code (a part of SWPC).

1) Character matrix generated by the combination of radical and phonetic component codes

- Figure below illustrates some Chinese characters in the form of a matrix. The Chinese character pictures in the first and third columns are the same as those in the third row. These characters have three identities.

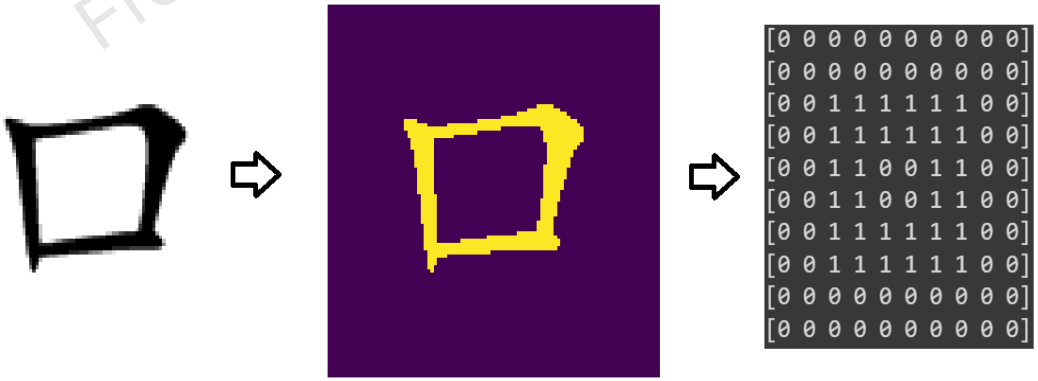
			Phonetic component (部件声码)			
Character root (字根编码)	Matrix (字/码矩阵)		351100	951100	533100	977100
			口	女	斤	又
口 mouth 35353535	Radical 偏旁 形码	口 3560	吕 name 3560351100		听 listen 3560533100	叹 sight 3560977100
女 woman 95959595		女 9575	如 as 9575351100	妓 quarrel 9575951100	妡 name 9575533100	奴 slave 9575977100
斤 catty 53515131		斤 5372			所 catties 5372533100	
又 again 97979797		9774 又				双 double 9774977100

2) Image/text multimodal processing algorithm

- With the complexity of $O(n^2)$, by combining SWPC with the image 0-1 matrix of Chinese characters.

Step 1: coding the root characters. The root characters are coded using the SWPC method and saved in a data set. Following the WB principle of character splitting, there are approximately 300 root characters.

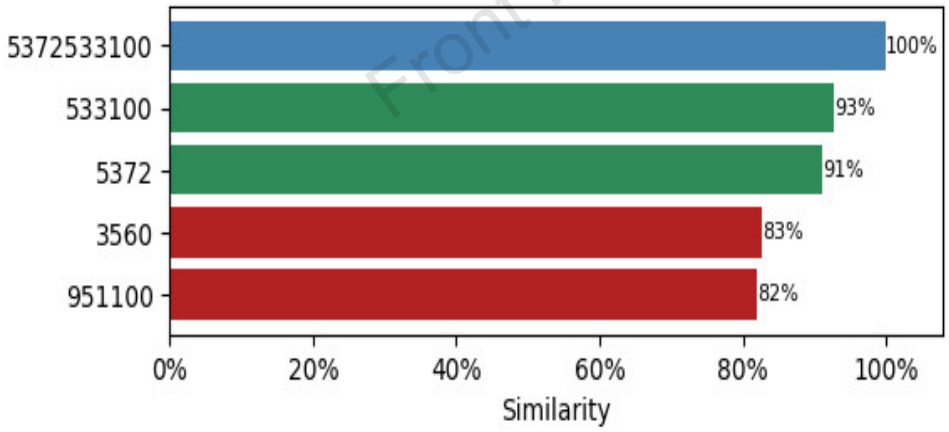
Step 2: pre-processing the images of the afore-mentioned root characters, including radicals and phonetic components. Each root (or radical or component) image is binarized using the Otsu threshold method (Otsu, 1979).



2) Image/text multimodal processing algorithm

Step 3: establishing a Chinese character root image/text (coding) database. After digitizing all the images based on the coded root characters and their SWPC (from step 1), a database of roots (radicals and phonetic components) of Chinese characters is created.

Step 4: similarity calculation between the image codes of Chinese characters (image matrix). We use the HM distance similarity (Hamming, 1950) to predict the similarity between image matrices of Chinese characters.



2) Image/text multimodal processing algorithm

Step 5: multimodal processing of Chinese character using image/text. Based on the root database of characters (with the image matrix and SWPC of radical and component), various tasks of multimodal processing can be performed. Four roots and their associated radicals and components are processed to generate nine Chinese characters.

Step 6: result generation. Given a Chinese character, with the above steps our algorithm can identify which combinations of radical and phonetic components generate the similar character through similarity analysis.

IV. SWPC for analogical reasoning

- Considering the pictophonetic features of Chinese characters and the Chinese language program, we divide the question pairs in the CA8 data set into two categories (Li S et al., 2018).
- In standard language scenarios, new characters or words are generated based on Chinese language regulations and customs. If we have a question pair (A:B; C:*X*), where A, B, and C are known characters or words, and *X* is an unknown, we can use SWPC to encode the characters or words. Hereinafter, unless otherwise specified, *f*(·)=SWPC(·). With *f*(A), *f*(B), and *f*(C), the generation of *f*(*X*) of character or word *X* can be expressed as follows:

$$f(X) = f(\mathrm{B}) - f(\mathrm{A}) + f(\mathrm{C}).$$

The encoding requirement in this equation ensures that the differences between characters or words are distinguishable by morphological rules.

SWPC for analogical reasoning – case 2

In non-standardized language scenarios, new words are generated with maximum similarity. Different from the aforementioned scenario, it becomes necessary to calculate the similarity between characters or words due to the flexibility and idiomatic nature of Chinese.

To account for uncertain language environments where exact but similar words may not exist, we can search for characters or words with the highest similarity to *f*(*X*)=SWPC(*X*) relative to the known A, B, and C. This can be expressed as follows:

$$\text{Find } X: \arg\max\{\text{Sim}(f(X), f(\text{B}) - f(\text{A}) + f(\text{C}))\}.$$

1) Analogical modes for CA8-Mor-10177

1. (A, A-P-A) mode

The CA8-Mor-10177 data subset presents several (A, A-一 (one)-A) patterns, such as 避 (avoid) (A) and 补 (make up) (*X*): (避, 避一避; 补, 补一补). The generation of a new word can be expressed as follows:

$$f(X\mathrm{P}X) = \mathrm{Concat}((f(\mathrm{AP}) - f(\mathrm{A}\mathbf{0}) + f(X\mathbf{0})), f(X)),$$

The correctness of Eq. (4) can be verified using the HM distance:

$$\begin{aligned}&\mathrm{HMDis}(f(\mathrm{A}), f(X))\\ &= [\mathrm{HMDis}(f(\mathrm{APA}), f(X\mathrm{P}X))]/2,\end{aligned}$$

where HMDis(·) represents the HM distance between the two strings of SWPC.

1) Analogical modes for CA8-Mor-10177

2. (A, A-P-A-Q) mode

The CA8-Mor-10177 data subset presents several questions: (说(A), 说(A)来(P)说(A)去(Q)); 比(*X*), 比(*X*)来(P)比(*X*)去(Q)), involving 说 (say), 来 (come), 去 (go), and 比 (compare). Therefore, for these Chinese word patterns, we propose a more general A-P-A-Q model, namely (A, A-P-A-Q; *X*, *X*-P-*X*-Q). Using SWPC, we have

$$f(X\mathrm{P}X\mathrm{Q}) = \mathrm{Concat}((f(\mathrm{AP}) - f(\mathrm{A}\mathbf{0}) + f(X\mathbf{0})), \\ (f(\mathrm{AQ}) - f(\mathrm{A}\mathbf{0}) + f(X\mathbf{0}))),$$

where the length of **0** equals that of the previous word code used to fill in the blanks.

1) Analogical modes for CA8-Mor-10177

3. (AB, A-P-AB) overlapping mode

The CA8-Mor-10177 data subset includes several questions, such as (慌张(AB), 慌(A), 里(P), 慌 张(AB); 马虎(*XY*); 马(*X*), 里(P)马虎(*XY*)), involving 慌张 (panic), 里 (connection), and 马虎 (careless). This mode can be summarized as (AB, A-P-AB; *XY*, *X*-P-*XY*). Like the previous model, we have

$$f(X\mathrm{P}XY) = \mathrm{Concat}((f(\mathrm{AP}) - f(\mathrm{A}\mathbf{0}) + f(X\mathbf{0})), f(XY)).$$

1) Analogical modes for CA8-Mor-10177

4. (AB, AA-BB) overlapping mode

Within the CA8-Mor-10177 data subset, there are several problems: (安全(AB), 安安(AA)全全(BB); 快乐(*XY*), 快快(*XX*)乐乐(*YY*)), involving 安全 (safety) and 快乐 (happy). The model can be summarized as follows: (AB, AA-BB; *XY*, *XX*-*YY*). Like the previous model, we have

$$f(XXYY) = f(\mathrm{AABB}) - f(\mathbf{0}\mathrm{AB}\mathbf{0}) + f(\mathbf{0}XY\mathbf{0}) - f(\mathrm{A}\mathbf{00}\mathrm{B}) + f(X\mathbf{00}\,Y).$$

2) Prefix word pattern

The CA8-Mor-10177 data subset includes 21 patterns of semi-prefixes, such as 大 (big), 小 (small), 老 (old), 第 (order), and 亚 (second), totaling 2553 questions. This pattern can be summarized as (A, PA; *X*, P*X*), e.g., 虎-老虎 (tiger-tiger) and 鹰-老 鹰 (eagle-eagle). If A or P represents a multi-character word, the pattern can be extended to (AB, PQ-AB; *XY*, PQ-*XY*), e.g., 草原-大草原 (grassland-prairie) and 都市-大都市 (city-metropolis). Like the above formulas, we have the model as follows:

$$f(\mathrm{PQ}XY) = f(\mathrm{PQAB}) - f(\mathbf{00}\mathrm{AB}) + f(\mathbf{00}XY).$$

3) Suffix word pattern

The CA8-Mor-10177 data subset contains 41 patterns of semi-suffixes, including 者 (zhe), 式 (shi), 性 (sex), and others, totaling 2535 questions, e.g., 我 (I, A), 我们 (we, AP); 你 (you, *X*), 你们 (you, *X*P). This pattern can be summarized as (A, AP; *X*, *X*P). If A or P represents a multi-character word, the pattern can be extended to (AB, AB-PQ; *XY*, *XY*-PQ), e.g., 乐观, 乐观主义 (optimism-optimism) and 悲观, 悲观主义 (pessimism-pessimism). Referring to the above formulas, we have

$$f(XY\mathrm{PQ}) = f(\mathrm{ABPQ}) - f(\mathrm{AB}\mathbf{00}) + f(XY\mathbf{00}).$$

4) Prompting by using SWPC

- We devised a Chinese Q&A prompt using the CA8-SEM-7636 data set (Li S et al., 2018) to assess the performance of LLMs. Use the following forms: (1) (*X*P, *Y*P; *X*Q, *Y*Q): (公狼male wolf, 母狼female wolf; 公熊male bear, 母熊female bear); (2) (*X*P, *Y*P; MR, FR): (公狼male wolf, 母狼female wolf; 雄鸟male bird, 雌鸟female bird); (3) to have the correct animal name (*X*Q, *Y*Q; MT, ?): (公熊, 母熊; 雄兔male rabbit, ?).
- Liu PF et al. (2023) proposed Eq. (11) to search over the set of potential answers "*z*" by calculating the probability of their corresponding filled prompts using a pre-trained language model $P(\cdot;\theta)$:

$$\hat{z} = \text{search}\ P(f_{\text{fill}}(x', z); \theta), \quad z \in \mathbb{Z},$$

where θ is the learning parameter from LLMs. Using word embedding to represent the characters or words (Mikolov et al., 2013), as in Eq. (3), we have

$$\text{Find}\ z : \text{argmaxSim}(f(z), f(\text{MR}) - f(\text{MQ}) + f(\text{FQ})),$$

Solving the analogical reasoning problem by prompt mode

All the Chinese characters (words) in the below table are represented by SWPC, which is used for the prompting and prediction processes. In this example, the Chinese program and habits are used to ensure reasoning (prompting) processing. In some cases, it is necessary to calculate the similarity. When applying the SWPC approach, $f(\cdot)$= SWPC(·). There is z=FR; i.e., in our example, F=雌 (female), R=兔 (rabbit).

Table 5 Solving the analogical reasoning problem by prompt mode*

Name	Notation	Example	Description
Input	x	(公熊, 母熊; 雄鸟, 雌鸟) ->	One or multiple texts
Output	y	(公熊, 母熊; 雄兔, ?)	Output other text
Prompting function	$f_{\text{prompt}}(x)$	(AP, BP; MQ, FQ) ->(AP, BP; [X], [Z})	A function that converts the input into a specific form by inserting the input x and adding a slot [Z], where z will be filled later
Prompt	x'	(AP, BP; MQ, FQ) ->(AP, BP; MR, [Z})	[X] is instantiated by input x' = MR but answer slot [Z] is not
Fill prompt	$f_{\text{fill}}(x',z)$	(MQ, FQ) ->(MR, ?R)	A prompt where slot [Z] is filled with any character related to R
Answered prompt	$f_{\text{fill}}(x',z^*)$	(MQ, FQ) ->(MR, FR)	A prompt where slot [Z] is filled with a true answer FR
Answer	z	FR (雌兔), XR (母兔), ...	To verify the Chinese program and customs

* Using a form similar to that in Liu PF et al. (2023)

V. Conclusions

- We propose the SWMP framework for Chinese language models. This framework integrates multi-modal information of Chinese characters, including pictophonetics, pinyin, images, and semantics, to enhance the effectiveness of CNLP.
- By conducting variability analysis of the pictophonetic coding of Chinese characters, such as FC numbers, and exploring similarity calculation methods, we propose to augment and normalize the WB numerical coding from 11–55 to 11–99.
- At the same time, the numerical coding method of pinyin is developed; i.e., the initial consonants are coded by GB uppercase letters/numbers and the vowels are coded by GB lowercase letters/numbers. Then they are augmented and normalized in the range of 11–99.

V. Conclusions

- Under the framework of SWMP, we introduce the concept of SWPC, which combines the expression of characters with Chinese grammar and flexible properties. With its moderate granularity of representation, SWPC possesses a generative and prompting mechanism for multimodal processing of Chinese characters.
- Considering the variability of numerical expressions of Chinese characters and the analysis of related similarity calculation methods, SWPC can effectively express the similarity between similar characters and the dissimilarity between characters.
- By combining SWPC with Chinese character image processing and other multimodal processing technologies, we propose different methods for Chinese character generation. Our research demonstrates how to establish a Chinese character data set including roots, radicals, and phonetic components, based on their SWPC, thus forming a Chinese character generative matrix that facilitates various CNLP tasks.

V. Conclusions

- By leveraging Chinese grammar, the phonological features of Chinese characters, and the requirements of language models, we establish analogical reasoning models for various word combinations using SWPC. As a result, the processing accuracy of word pairs, including repetition (AABB), prefix (PQAB), and suffix (ABPQ), can achieve 100% accuracy. These models are also used for the purpose of prompting Q&A Chinese language models.
- The application of SWPC to data sets like CA8 (Li S et al., 2018) enables high-precision analogical reasoning for word pairs conforming to Chinese grammar and pictophonetic properties.
- SWPC is applied in analyzing the COS960 data set (Huang JJ et al., 2019) to evaluate the strengths and weaknesses of various similarity calculation methods.

VI. Future work

- The SWMP framework we propose is just the first step in improving Chinese language models using the concept of Six-Writings. It still has some shortcomings that need to be addressed. For example, to enable multimodal processing of pictophonetic, pinyin, image, property, audio/video, and understanding, we need to integrate SWMP into the language model, or even establish a new Chinese language model.
- We have presented only the theory and methodology of SWMP/SWPC and not addressed the establishment of a Chinese character database (e.g., 3500 common Chinese characters), which would facilitate SWPC/image coding based on roots, radicals, and components.
- It is necessary to establish a common Chinese character root database to realize the coding of roots, radicals, and components by SWPC with the support from the academic community and even the government.

Li WEIGANG, corresponding author of this paper, is a professor and coordinator of TransLab of the Department of Computer Science at the University of Brasilia, Brazil. He received his PhD from the Aeronautics Institute of Technology (ITA), Brazil, in 1994. He coordinated various research projects from CAPES, CNPq, FINEP, FAPESP, Atech, and Boeing, and advised more than 120 students including PhD and post-doctoral researchers. His research interests include artificial intelligence in air traffic management and natural language processing (https://orcid.org/0000-0003-1826-1850).

Mayara Chew MARINHO is a data scientist and master student of the Department of Computer Science at the University of Brasilia (UnB), Brazil. She received her BS degree in computer engineering from the University of Brasilia (UnB) in 2023. Her research area is machine learning with focus on visualization techniques (https://orcid.org/0009-0004-3159-7804).

Denise Leyi LI received her PhD in economic theory from the University of São Paulo (USP) in 2021. She is part of the Regional and Urban Economics Lab (NEREUS) in USP. Her research interests include application of quantitative methods to the areas of consumer demand, education and economic evaluation of public policies, and also computational applications (https://orcid.org/0000-0003-1826-1850).

Vitor Vasconcelos DE OLIVEIRA is a data scientist and master student of the Department of Computer Science at the University of Brasilia (UnB), Brazil. He received his BS degree in computer science from the University of Brasilia (UnB) in 2023. His research area is machine learning with focus on natural language processing (https://orcid.org/0009-0001-0026-9240).

References

[1] Xu S, 1997. Discussing Writing and Explaining Characters. Yuelu Publishing House, Changsha, China (in Chinese).

[2] Zhang B, 2008. Newly Edited Chinese Language (2nd Ed.). Fudan University Publishing, Shanghai, China (in Chinese).

[3] Yeromiyan T, 2023. The Six Types of Chinese Characters. https://studycli.org/chinese-characters/types-of-chinese-characters/ [Accessed on May 30, 2023].

[4] Otsu N, 1979. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern, 9(1):62-66. https://doi.org/10.1109/TSMC.1979.4310076

[5] Mikolov T, Yih WT, Zweig G, 2013. Linguistic regularities in continuous space word representations. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.746-751.

[6] Li S, Zhao Z, Hu RF, et al., 2018. Analogical reasoning on Chinese morphological and semantic relations. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.138-143.

[7] Huang JJ, Qi FC, Yang CH, et al., 2019. COS960: a Chinese word similarity dataset of 960 word pairs. https://arxiv.org/abs/1906.00247

[8] Liu PF, Yuan WZ, Fu JL, et al., 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv, 55(9):195. https://doi.org/10.1145/3560815

[9] Zhou J, Ke P, Qiu XP, et al., 2023. ChatGPT: potential, prospects, and limitations. Front Inform Technol Electron Eng, early access. https://doi.org/10.1631/FITEE.2300089