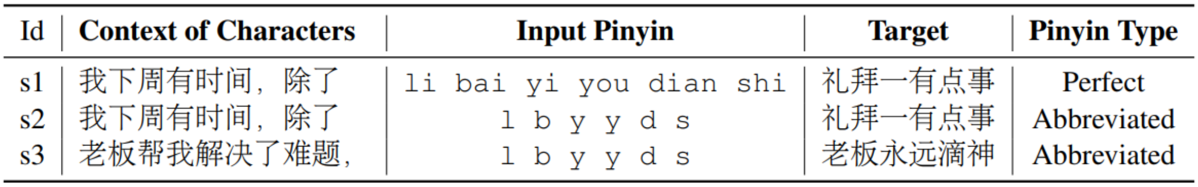
Kiara Madeam
Professor Mazidi
Paper: https://aclanthology.org/2022.acl-long.133.pdf
Authors and Affiliations:
(1) Minghuan Tan: Singapore Management University
(2) Yong Dai: Tencent AI Lab
(3) Duyu Tang: Tencent AI Lab
(4) Zhangyin Feng: Tencent AI Lab
(5) Guoping Huang: Tencent AI Lab
(6) Jing Jiang: Singapore Management University
(7) Jiwei Li: Zhejiang University
(8) Shuming Shi: Tencent AI Lab

Chinese GPT and Pinyin Input

In *Exploring and Adapting Chinese GPT to Pinyin Input Method,* Tan et. al. (2021) address the challenge of leveraging Chinese GPT to handle Pinyin input methods. Pinyin is a standard system of romanized spelling that allows users to input Chinese text using a standard keyboard. Although GPT has been developed successfully for Chinese and many other languages, existing Chinese language models are not optimized for this type of input despite it being utilized by hundreds of millions of Chinese speakers on their computers and cellphones. This research team is the first to explore the use of Chinese GPT for pinyin input methods. Key findings from the research show that frozen GPT achieves amazing results on perfect pinyin but results dramatically worsen with abbreviated pinyin. Context enrichment and pinyin-constrained training improve the performance of frozen GPT with abbreviated pinyin, and as the context of Chinese characters becomes longer, GPT-based models show an improvement in performance.

A Generative Pre-Trained Transformer (GPT) is a language model based on the Transformer architecture, and can be further tweaked to carry out more specific language processing tasks. Previous research has shown that GPT has the potential to adapt to different input methods, such as calligraphy or speech recognition, by making changes to training strategies or data preprocessing. Pinyin keyboards allow for pronunciation-based entering of Chinese characters using the Roman alphabet and for a given pinyin, the input system will return a list of Chinese characters with that pronunciation for the user to choose from. Each pinyin is made up of initials and finals, and in most cases, a Chinese character is spelled with one initial followed by one final. People may also choose to enter either “perfect pinyin,” where initials and finals of all characters are entered, or “abbreviated pinyin,” where only the initials of characters are entered. A large contributor to the problem of using pinyin input methods is that an abbreviated pinyin maps to many perfect pinyin, revealing an exponentially larger pool of Chinese characters that the pinyin could be referring to. In one language model, the researchers explored concatenation of supplementary context to the tokenization and positioning scheme, and in another language model a pinyin-embedding layer was added at the bottom of GPT. After enriching input and training the language models with pinyin-constrained vocabulary, the ability to distinguish between characters pronounced with the same pinyin dramatically improves.



Id	Context of Characters	Input Pinyin	Target	Pinyin Type
s1	我下周有时间，除了	`li bai yi you dian shi`	礼拜一有点事	Perfect
s2	我下周有时间，除了	`l b y y d s`	礼拜一有点事	Abbreviated
s3	老板帮我解决了难题，	`l b y y d s`	老板永远滴神	Abbreviated

Table 2: Illustrative examples of the task of pinyin input method with perfect pinyin and abbreviated pinyin. In s3, the input pinyin “`l b y y d s`” is the abbreviation of “`lao ban yong yuan di shen`”. The translations of s1 and s3 are “I am free next week except for the next Monday.” and “Boss helps me overcome the obstacle. You are the greatest of all time.”, respectively.

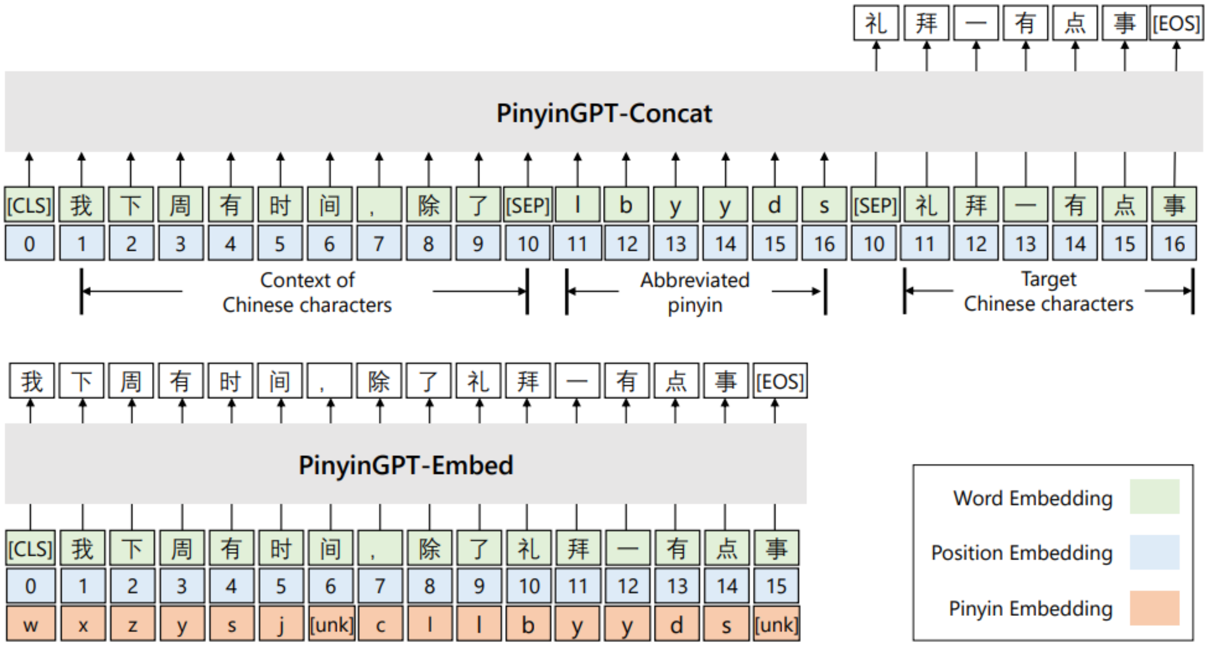

礼 拜 一 有 点 事 [EOS]
PinyinGPT-Concat
[CLS] 我 下 周 有 时 间 ， 除 了 [SEP] l b y y d s [SEP] 礼 拜 一 有 点 事
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 10 11 12 13 14 15 16
Context of Chinese characters
Abbreviated pinyin
Target Chinese characters
我 下 周 有 时 间 ， 除 了 礼 拜 一 有 点 事 [EOS]
PinyinGPT-Embed
[CLS] 我 下 周 有 时 间 ， 除 了 礼 拜 一 有 点 事
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
w x z y s j [unk] c l l b y y d s [unk]
Word Embedding
Position Embedding
Pinyin Embedding

Figure 1: An illustration of the training process of Pinyin-Concat (top) and Pinyin-Embed (bottom), respectively. The example is same as the instance of s2 from Table 2.

The researchers made use of benchmark datasets like *People's Daily* and created their own dataset from 822 million web pages called the *WuDaoCorpora.* They compared their results with GPT baselines, such as Google IME, On-OMWA, and On-P2C, and were able to achieve comparative performance using perfect pinyin but still experienced drastic drops in performance for abbreviated pinyin. Adding pinyin context horizontally with concatenation achieved better results than vertically embedding a pinyin layer and fixed GPT parameters were not as promising as fine tuned parameters. An ablation study was performed to understand the importance of each component proposed by the researchers and through their model analysis, the team concluded that both strategies are responsible for performance boosts.

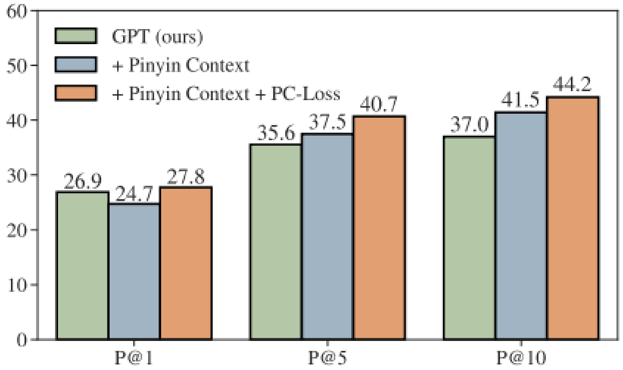


Figure 2: Ablation study for concatenating pinyin context and pinyin-constrained training.

This study opens up many doors for future research, such as extending to phonetic input methods, and training their adapted GPT model for more complex tasks. This research represents a significant contribution to the accessibility of language processing systems for Chinese speakers using pinyin input and generalized improvements in text classification and other NLP tasks across the board. The insights revealed in this study can also inform future studies on the challenges and solutions for similar tasks in other languages.