What's behind the MASK?
— A Chinese Idiom Cloze Test Using Fine-tuning and Prompt Engineering

Dave (Hao) Jia, Jessie (Jiexin) Kuang, Grace (Mengyuan) Qiao,
Sherry (Ziyun) Wang, Liangchen Xia,

COLX 585: Trends in Computational Linguistics

Dr. Jian Zhu

April 28th, 2023

Abstract
Chinese idioms, as a unique language phenomenon, pose challenges for existing language models due to their cultural and contextual complexity. In this project report, we present a comprehensive study of Chinese idioms, focusing on the development and evaluation of a large-scale Chinese cloze test dataset, ChID. We explore the effectiveness of various transformer models, including BERT, T5, T5-small, mT5-small, GPT-2, and ChatGPT, and compare their performance when explicitly fine-tuned for this task against their performance when prompted to perform the same task without fine-tuning. We found that the GPT-2 model works best with both fine-tuning and prompt engineering. Our results will provide insights into the most effective strategies for leveraging large language models (LLMs) for Chinese idiom comprehension through fine-tuning or carefully designed prompting, and contribute to a deeper understanding of the challenges and opportunities in using transformer models for this task.

---

1. Introduction

1.1 Background Concept: Chengyu

In recent years, machine reading comprehension and completion have been greatly improved by a variety of corpora and task settings. Idioms, or "成语" (chengyu) in Chinese, are often used to succinctly summarize or paraphrase text and typically consist of four characters. They are a particularly interesting linguistic phenomenon in the Chinese language. This work is part of a larger effort to develop datasets that address different language phenomena in English. Compared to other types of words, idioms are unique in that they often have non-compositional and metaphorical meanings, making it important to develop effective representations of them. Additionally, the presence of near-synonyms, or words with similar but not identical meanings, can pose a challenge for machines when choosing the correct idiom for a given context. As idioms are widely used in daily communication and literature, it is a new challenge to assess the ability of machines to understand and represent idioms in Chinese reading comprehension. For example, an easy example of a Chinese idiom present in **Figure 1** is written as “冰山一角”. Its English translation is one corner of an ice mountain, while its Chinese meaning is only a small portion of a larger issue or problem is visible, while the majority of the issue or problem remains hidden or unknown. This is an example in which the Chinese idiom implicates more meaning than the words.

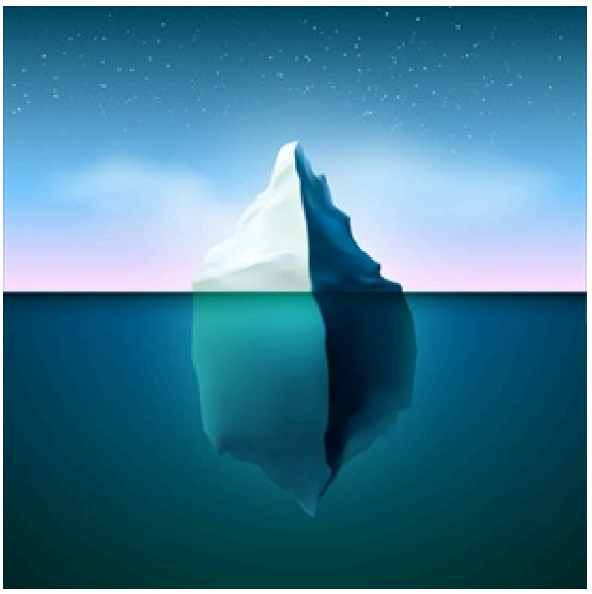

Figure 1: Chinese idiom example: “冰山一角”(bīng shān yī jiǎo).

Another example is “亡羊补牢” from **Figure 2**. The English translation is to mend the fence after sheep are lost while this idiom has a metaphorical meaning: never be late to try. Thus, understanding and representing this idiom may require an understanding of its corresponding cultural history. Furthermore, even compositional idioms such as "亡羊补牢" are likely to have multiple meanings due to the polysemy of individual characters, which makes the representation of idioms a challenging problem.

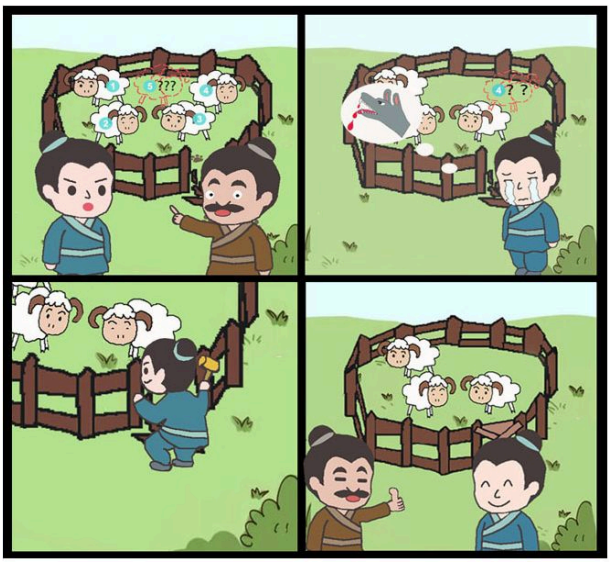

Figure 2: Chinese idiom example: “亡羊补牢” (wáng yáng bǔ láo).

1.2 Literature Review/Related Work

Cloze test is a typical reading comprehension task that is essential to assessing machine reading ability. Researchers have created a number of cloze-style reading comprehension datasets to facilitate cloze research. Chinese idioms and English slang are special forms of language with unique expressions and a long history. The non-literal deep metaphor of Chinese idioms has brought great challenges to the research on Chinese idiom

cloze task. ChID has settings similar to CLOTH, where the answers are selected from the given options. But unlike most existing cloze test corpora, the answers from ChID usually do not appear in the context.

Zheng et al. proposed the Chinese idiom cloze task and used Bi-LSTM, Attentive Reader, and Standford Attentive Reader as benchmark models. Attentive Reader added an attention mechanism to Bi-LSTM, and Stanford Attentive Reader adopted a bilinear function as a matching function to calculate attention weight. With the emergence of the BERT model, researchers have gradually adopted the BERT model to solve the Chinese idiom Cloze task. The first BERT-based dual-embedded idiom cloze model learned the dual-embedding of idioms to predict the idioms in the blanks. However, the basic BERT models still face challenges in dealing with long sequences and understanding the metaphorical meanings of idioms. Subsequently, by introducing external knowledge, idiom definitions and idioms' characters can be used to correct the misuse of idioms. Long et al. found that the literal meaning of many idioms was significantly different from their metaphorical meaning, so they constructed a synonym graph according to the meanings of idioms and encoded the idiom into a new representation. However, it is more challenging for this method to distinguish idioms with similar meanings.

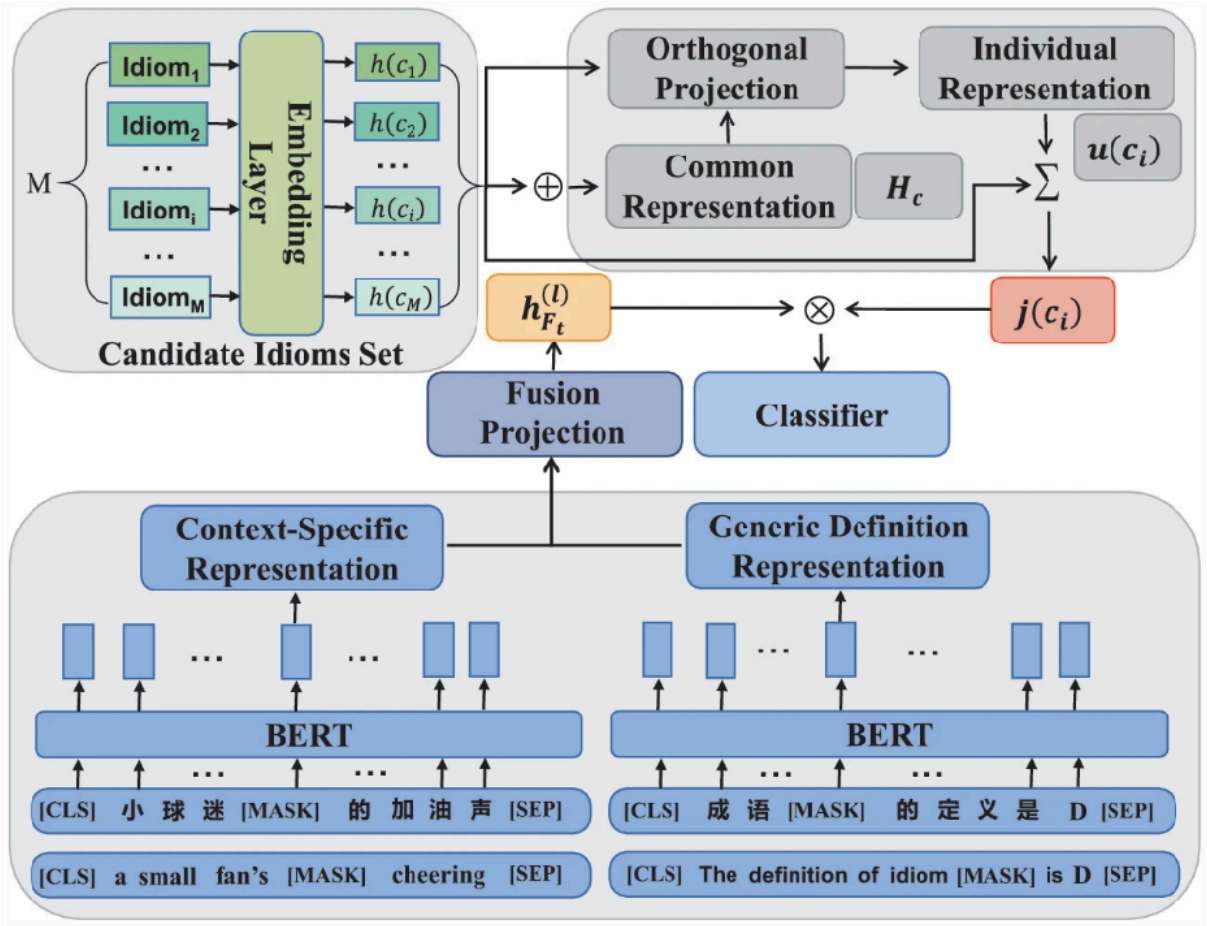

Figure 2: Model structure (Ying et al., 2023).

1.3 Motivation

As students of linguistics, we recognize that semantic analysis in the Chinese language corpus can be particularly difficult due to factors such as ambiguity, polysemy, and the need for contextual information. Chinese boasts a vast vocabulary and many near-synonyms, which can pose challenges for accurate analysis and comprehension of meaning. Moreover, most transformer-based models are currently trained on English corpus, which presents a constraint for semantic analysis in Chinese. To overcome this limitation, we can explore bilingual tasks that involve both English and Chinese languages. This approach allows us to leverage the unique strengths of each language and gain a more nuanced understanding of meaning across different cultural and contextual boundaries.

In summary, the difficulties inherent in semantic analysis for Chinese, in conjunction with the limitations of transformer-based models trained solely on English corpus, emphasize the importance of investigating bilingual tasks to enhance our comprehension of meaning across different languages and cultures.

2. Datasets

Our proposed dataset for cloze tests, ChID, is a comprehensive collection of Chinese idioms. It includes 581k passages and 729k blank spaces, covering a wide range of domains. In ChID, each passage's idioms are replaced with blank symbols, *#idiom#*, and a list of potential idioms, including the correct one, is provided for each blank (see **Figure 4**).

Passage & Blanks		可是有一个时期大家*#idiom-0#*，不大敢露面， 只有她一个人倚在阳台上看排队的兵走过。 However, there was a period when everyone *#idiom-0#* and was scared to show up. Only she leaned on the balcony and watched the soldiers passing by.	
#idiom-0# **options**	**Correct**	深居简出	Be unwilling to contact people
	Similar	销声匿迹	To disappear from the scene
		离群索居	To stay away from the crowd and live alone
		安分守己	To know one's place
	Random	一帆风顺	To proceed smoothly without a hitch
		文不对题	Be irrelevant to the subject
		万里长征	A long and arduous undertaking

Figure 4: An example in ChID. Each piece of data comprises a passage with multiple blanks that replace the original idioms. In the given example, there is only one blank, and several options are provided for each blank. The options consist of a correct answer, three similar idioms, and three random ones (Zheng et al., 2019).

The dataset used in this study was obtained from a GitHub repository (https://github.com/chujiezheng/ChID-Dataset). The dataset is structured as shown in **Figure 5** and includes four key components: the content, which is the given passage where the original idioms are replaced by placeholders (#idiom#); the realCount, which is the number of placeholders or blanks in the passage; the groundTruth, which provides the correct answers

for each placeholder in the order they appear in the passage; and the candidates, which lists the potential answers for each placeholder in the same order as the groundTruth.

```
dataset = ["""{
    "content": "世锦赛的整体水平远高于亚洲杯，要如同亚洲杯那样“鱼与熊掌兼得”，就需要各方面密切配合、#idiom#。作为主帅的俞觉敏，除了得打破保守思想，敢于破格用人，还得巧于用兵、#idiom#、灵活排阵，指挥得当，力争通过比赛推新人、出佳绩、出新的战斗力。",
    "realCount": 2,
    "groundTruth": ["通力合作", "有的放矢"],
    "candidates": [
        ["凭空捏造", "高头大马", "通力合作", "同舟共济", "和衷共济", "蓬头垢面", "紧锣密鼓"],
        ["叫苦连天", "量体裁衣", "金榜题名", "百战不殆", "知彼知己", "有的放矢", "风流才子"]
    ]
}"""]
```

Figure 5: An example of the dataset.

To prepare the ChID dataset for fine-tuning and prompt engineering, we performed preprocessing steps. The dataset was loaded from text files and split into separate training and validation sets. To expedite model training, the original dataset was sliced to create a smaller training set, consisting of 20,000 samples. We then evaluated our models on a validation and test set, which contained 3,000 samples. This approach was necessary as the large size of the original dataset would have taken a significant amount of time to process on all models.

3. Methods

3.1 Fine-tuning

We fine-tuned BERT, T5, T5-small, mT5-small, and GPT-2 models.

Bert is a multilingual model, which means there is no need to find a Chinese pre-trained model to fine-tune. However, as we realized a pre-trained Chinese model might help us get a higher score, we found bert-base-chinese from HuggingFace (https://huggingface.co/bert-base-chinese). This model has been pre-trained for Chinese, and training and random input masking have been applied independently to word pieces. We also use BertTokenizer as the tokenizer which the model can understand input word by word. The BERT-Base Chinese model consists of 12 transformer layers, with a hidden size of 768 and 12 self-attention heads. Layer normalization and residual connections are used between each layer to improve training stability and performance. The model uses the ReLU activation function. For regularization during training, the model employs dropout with a rate of 0.1. The cost function used during pretraining is the masked language modelling (MLM) objective. Additionally, the model is also trained using the next sentence prediction (NSP) objective, which aims to predict whether two sentences in a sequence are consecutive or not. The model is optimized using the Adam optimization algorithm with a learning rate of 1e-4.

To fine-tune the T5 model, we utilized the AutoTokenizer to tokenize input sentences and labels. A custom collate function was created to ensure the input and labels were of the same length and attention masks were generated for both. The fine-tuning was done on a pre-trained T5 model available in the Hugging Face Transformers library, which is an AutoModelForSeq2SeqLM. This model has 12 layers, 768 hidden units, and 12 attention heads. For training, we used a dataset of 20,000 examples for 5 epochs and employed the Adam optimizer with a learning rate of 2e-5 and a batch size of 8. During training, the cross-entropy loss function was used to calculate the model's loss.

We use T5-small because we want to see how the smaller model performs for the similar task. mT5-small was used to (roughly) see if the single Chinese language model can perform better than the multilingual language model. After preprocessing the data and splitting the data into training and validation sets, we tokenized the input sentences and labels using the BertTokenizer to feed in the T5-small model, because it is pre-trained on Chinese to predict the missing single character. For more information, please refer to uer/t5-small-chinese-cluecorpussmall (https://huggingface.co/uer/t5-small-chinese-cluecorpussmall). This model has 6 layers and 512 hidden units. We tokenized the input sentences and labels using the AutoTokenizer to feed in the T5-small model. For more information, please refer to google/mt5-small (https://huggingface.co/google/mt5-small). We used a custom collate function to pad the input and labels to the same length and to create attention masks for the input and labels. We fine-tuned these models on our Chinese idiom completion task by training it on a dataset of 20,000 examples for 5 epochs. During training, we used the Adam optimizer with a learning rate of 5e-5 and a batch size of 8. We used the cross-entropy loss function to compute the model's loss.

For GPT-2, we used the "uer/GPT-2-chinese-cluecorpussmall" pre-trained model available in the HuggingFace library (https://huggingface.co/uer/GPT-2-chinese-cluecorpussmall) which was specifically designed for Chinese text generation. The GPT-2 model used in this project has 12 transformer layers, 768 hidden units, and 12 attention heads. It employs layer normalization and residual connections between each layer. The model has a vocabulary size of 30,000 subwords and uses dropout with a rate of 0.1 as its regularization technique during training. The activation function used in the model is GELU. For pretraining, the model was trained on the Chinese ClueCorpus dataset using a masked language modelling (MLM) objective and the Adam optimization algorithm with a learning rate of 2e-5. After loading the GPT-2 model, we selected the BertTokenizer, which tokenizes text into subwords and includes special tokens for the start and end of sentence markers as well as for separating multiple sentences. We used a seq2seq model for fine-tuning our model. We trained our model on a subset of 20,000 training examples and split 2,000 examples each for the validation and test sets. For evaluation, we decided that all of our large language models have epochs of 5. For each epoch, we printed the Eval loss value and Eval Acc.

3.2 Prompt Engineering

We tried prompt engineering on models: T5, T5-small, mT5-small, GPT-2 and ChatGPT.

In order to prepare the dataset for prompt engineering, we conducted data preprocessing by replacing the mask in the content with the candidate idioms. During the model evaluation phase, we employed several different approaches: ranking the most appropriate idiom to fill in the blanks in the sentence (see **Figure 6**) or ranking the most appropriate idiom from among the candidates (see **Figure 7**) or providing the list of idiom candidates first and let the machine input them into the sentence (see **Figure 8**).

```
Input example:
 [CLS] 请 从 （ ） 里 选 择 出 最 合 适 的 成 语 ： （ 超 凡 入 圣 | 骨 瘦 如 柴 | 青 面 獠 牙 | 虎 背 熊 腰 | 成 人 之 美 | 肥 头 大 耳
| 神 不 守 舍 ） 的 掌 柜 只 穿 一 件 衬 衫 ， 坐 在 柜 台 里 。 几 个 堂 倌 穿 着 脏 得 发 黑 的 白 工 作 服 ， 因 为 没 有 顾 客 ， 都 散
坐 在 桌 子 旁 。 这 当 儿 看 到 这 位 不 寻 常 的 客 人 ， 都 露 出 好 奇 的 神 色 列 宁 曾 批 评 他 理 论 上 的 错 误 ， 同 时 认 为 他
[UNK] 所 写 的 全 部 哲 学 ， 赶 紧 迎 上 前 来 伺 候 。 聂 赫 留 朵 夫 要 了 一 瓶 矿 泉 水 ， 在 离 窗 较 远 的 地 方 挨 着 一 张 铺 有
肮 脏 桌 布 的 小 桌 坐 下 。 [SEP] 肥 头 大 耳 [SEP]
```

Figure 6: An example of processed data for prompt engineering. Coded as prompt1.

```
Input example:
 [CLS] 请 依 次 从 （ 超 凡 入 圣 | 骨 瘦 如 柴 | 青 面 獠 牙 | 虎 背 熊 腰 | 成 人 之 美 | 肥 头 大 耳 | 神 不 守 舍 ） 选 择 出 最 合
适 的 成 语 填 入 _ ： _ 的 掌 柜 只 穿 一 件 衬 衫 ， 坐 在 柜 台 里 。 几 个 堂 倌 穿 着 脏 得 发 黑 的 白 工 作 服 ， 因 为 没 有 顾 客
， 都 散 坐 在 桌 子 旁 。 这 当 儿 看 到 这 位 不 寻 常 的 客 人 ， 都 露 出 好 奇 的 神 色 列 宁 曾 批 评 他 理 论 上 的 错 误 ， 同 时
认 为 他 [UNK] 所 写 的 全 部 哲 学 ， 赶 紧 迎 上 前 来 伺 候 。 聂 赫 留 朵 夫 要 了 一 瓶 矿 泉 水 ， 在 离 窗 较 远 的 地 方 挨 着 一
张 铺 有 肮 脏 桌 布 的 小 桌 坐 下 。 [SEP] 肥 头 大 耳 [SEP]
```

Figure 7: An example of processed data for prompt engineering. Coded as prompt2.

```
Input example:
 [CLS] 选 择 ： [ [ [UNK] 凭 空 捏 造 [UNK], [UNK] 高 头 大 马 [UNK], [UNK] 通 力 合 作 [UNK], [UNK] 同 舟 共 济 [UNK], [UNK] 和
衷 共 济 [UNK], [UNK] 蓬 头 垢 面 [UNK], [UNK] 紧 锣 密 鼓 [UNK] ], [ [UNK] 叫 苦 连 天 [UNK], [UNK] 量 体 裁 衣 [UNK], [UNK] 金
榜 题 名 [UNK], [UNK] 百 战 不 殆 [UNK], [UNK] 知 彼 知 己 [UNK], [UNK] 有 的 放 矢 [UNK], [UNK] 风 流 才 子 [UNK] ] ] 输 入 ： [UN
K] 世 锦 赛 的 整 体 水 平 远 高 于 亚 洲 杯 ， 要 如 同 亚 洲 杯 那 样 [UNK] 鱼 与 熊 掌 兼 得 [UNK] ， 就 需 要 各 方 面 密 切 配 合 、
# idiom # 。 作 为 主 帅 的 俞 觉 敏 ， 除 了 得 打 破 保 守 思 想 ， 敢 于 破 格 用 人 ， 还 得 巧 于 用 兵 、 # idiom # 、 灵 活 排 阵
， 指 挥 得 当 ， 力 争 通 过 比 赛 推 新 人 、 出 佳 绩 、 出 新 的 战 斗 力 。 [UNK] 输 出 ： 通 力 合 作, 有 的 放 矢 选 择 ： [ ['超
凡 入 圣 ','骨 瘦 如 柴 ','青 面 獠 牙 ','虎 背 熊 腰 ','成 人 之 美 ','肥 头 大 耳 ','神 不 守 舍'] ] 输 入 ： # idiom # 的 掌 柜 只
穿 一 件 衬 衫 ， 坐 在 柜 台 里 。 几 个 堂 倌 穿 着 脏 得 发 黑 的 白 工 作 服 ， 因 为 没 有 顾 客 ， 都 散 坐 在 桌 子 旁 。 这 当 儿
看 到 这 位 不 寻 常 的 客 人 ， 都 露 出 好 奇 的 神 色 列 宁 曾 批 评 他 理 论 上 的 错 误 ， 同 时 认 为 他 [UNK] 所 写 的 全 部 哲 学
， 赶 紧 迎 上 前 来 伺 候 。 聂 赫 留 朵 夫 要 了 一 瓶 矿 泉 水 ， 在 离 窗 较 远 的 地 方 挨 着 一 张 铺 有 肮 脏 桌 布 的 小 桌 坐 下
。 输 出 ： [SEP] 肥 头 大 耳 [SEP]
```

Figure 8: An example of processed data for prompt engineering. Coded as prompt3.

To preprocess the data, we tested three different prompts for all the models and determined that prompt2 was the best performer on the majority of the models. To be more specific about the data processing, prompt2 takes in three arguments from the data: text, candidates, and choice. The text argument is a string with a placeholder (#idiom#). The candidates' argument is a list of lists, with each inner list containing the candidate idioms that could fill in the corresponding placeholder in the text. The choice argument is an integer that specifies the index of the correct idiom. Prompt2 replaces the placeholder in the text with a blank underscore and constructs a new string with parentheses containing the candidate idioms separated by a vertical bar (|). The for loop iterates through the list and replaces each placeholder with the corresponding set of parentheses containing the candidate idioms. Finally, we remove the remaining parentheses from the string and return the resulting prompt string. We used this prompt to preprocess the data, ignoring any input texts with a length greater than 500 to avoid overloading the CPU or GPU.

It’s worth mentioning that ChatGPT is a proprietary model that is not publicly available for fine-tuning. As a result, it was not possible to fine-tune ChatGPT for the specific task of Chinese idiom comprehension. Instead, we evaluated ChatGPT's performance on the prompt engineering task and compared it to the performance of other transformer models that were fine-tuned for the task.

We used the OpenAI API gpt-3.5-turbo to sent the request to ChatGPT and get responses. The structure of our final prompts is:

- Two examples of choosing correct idiom(s) from the candidates given the paragraphs: One example included only 1 missing idiom from the paragraph, and another one include multiple missing idioms;

- The instruction which means “please follow the examples above and choose appropriate idiom(s) from the following brackets to fill in ‘#idiom#’”;
- The instruction which means “please only reply idioms, do not reply other characters”;
- The candidates in brackets: Each group of the candidates of the same missing idiom is in the same bracket; and
- The paragraph with the missing idiom(s)

We tried to use 3,000 test data to evaluate. However, the fee was used up halfway. It was too time consuming and money consuming to restart and collect all the responses, so we decided to use what we had. We got 2,096 responses. The amount is not perfect, but okay to represent the performance on 3,000 test data. We also did some minor cleaning on the response to clean up the mess signals.

4. Experiments

The models were evaluated using the f1-score, which is a harmonic mean of precision and recall, it is less sensitive to class imbalance than other metrics like accuracy. We separate an idiom into different tokens when training and we want to get the correct idiom in which none of the tokens in the idiom is wrong. When there a token from a predicted idiom is wrong, the f1-score would be low while we can still get high accuracy. Therefore, the f1-score would be better to describe the models' performance. We also used accuracy, which is an appropriate metric for our task as our goal is to choose the correct idiom from a list of candidates, and accuracy directly measures the proportion of correct predictions made by the model.

4.1 Fine-tuning

To evaluate our fine-tuned T5 model, we utilized a validation set consisting of 3,000 examples. We measured the performance of our model using accuracy and loss metrics and compared it to a baseline model that utilized a simple LSTM-based architecture. Our fine-tuned T5 model achieved an F1-score of 58.1% and a correct Chengyu of 2042 on the test set, representing a significant improvement over the baseline LSTM model, which achieved an F1-score of 8.4% and a correct Chengyu of 437 on the same test set. We observed that our fine-tuned T5 model was able to generalize well to new examples and did not overfit the training data.

We evaluated our fine-tuned T5-small model on a validation set of 3,000 examples. We measured the model's performance mainly using accuracy. Our fine-tuned T5-small model achieved an accuracy of 40.70% on the validation set and 41.38% on the test set. Our fine-tuned mT5-small model achieved an accuracy of 43.32% on the validation set and 42.34% on the test set, which is better than T5-small pre-trained on Chinese only. The reason can be that the multilingual models perform better than unilingual models, but the evidence is not enough. It can also be caused by the reason of data quality, pre-training methods, etc. We tried to use 200,000 data to train the T5-small model, and it performs much better, with an accuracy of 70.04% on the validation set and 67.89% on the test set. It took around 5 hours to train the model. Therefore, we decided to control the training time and still use 20,000 as our

training set scale. However, this shows that more data can make the effectiveness of fine tuning much better, at least on the T5-small model. In terms of batch size, we tried batch size 4 for both models, it was too slow and did not perform better. We also tried batch size 16, and even if we trained for more epochs, the performance of both models were worse. So we finally used batch size 8 to train both models.

4.2 Prompt Engineering

For T5-small and mT5-small, we tried to only use prompt engineering before fine-tuning, but both models got the results of 0% accuracy. It indicates that doing prompt engineering before fine-tuning makes no sense. We also tried to use different instructions that have the similar meaning with the fine-tuning instruction to instruct the model to choose the proper idiom from the candidates, in order to see if the model can really understand the instruction. However, the accuracy became only 5.98% and 16.24% for T5-small and mT5-small, respectively. The output also showed that although the model can make some correct predictions, they do not really understand the instruction to choose the proper idiom.

For both T5 and GPT-2, we ran our model on the three prompts (**Figure 6, 7 and 8**) and both models achieved highest score on prompt2, with f1-score of 8.4% and 63.8%, respectively.

For ChatGPT, we tried once on only 60 requests, and in each prompt we only include one example, and did not enforce not to return other characters other than the idiom(s). The performance was okay of around 24%, but it tended to return the whole paragraph. Therefore, we revised the prompts. The data amount 60 is also too small, and its accuracy may not be accurate, so we wanted to try the whole test set. As mentioned in the previous section "Method", we eventually did not get 3,000 but 2,096 instead.

5. Results

		BERT	T5	T5-small	mT5-small	GPT-2	ChatGPT
Fine-tuning	f1-score	0.544	0.581	0.415	0.424	0.644	-
	Accuracy	1722	2042	0.414	0.423	2179	-
Prompt Engineering	f1-score	-	0.084	0	0	0.638	0.203
	Accuracy	-	437	0	0	2166	0.188

Table 1: Table of evaluation of all models.

*Note: The results are slightly different from those in the presentation because all the models were retrained on 20,000 data.

*Note 2: The reason that the accuracy is written in numbers instead of percentages is that as every model uses different tokenizers, the total number of idioms from all the input sentences is different. The range of the total number of idioms from processing the test set with 3,000 sentences is between 2,800 and 3,000.

Based on the results (**Table 1**) obtained from the fine-tuning of BERT, T5, T5-small, mT5-small, and GPT-2 models on the Chinese idiom completion task, GPT-2 achieved the highest accuracy with 2179 correct predictions out of 3000 examples. The f1-score for GPT-2 was also comparatively high at 0.644. BERT and T5 models also performed relatively well with f1-scores of 0.544 and 0.581 respectively, but had lower accuracy scores. On the other hand, the smaller T5 and mT5 models had lower f1-scores and accuracy scores.

Based on the results of prompt engineering, GPT-2 has achieved the highest scores in terms of both accuracy and F1-score, with scores of 0.638. ChatGPT, T5, T5-small, and mT5-small, on the other hand, have scored significantly lower, with F1-scores of 0.203, 0.084, 0, and 0, respectively, and no accuracy scores for T5-small and mT5-small.

By comparing the scores on each model, we noticed a huge improvement of fine-tuning on T5, T5-small and mT5-small models. The performance of GPT-2 on fine-tuning and prompt engineering tasks are similar.

6. Conclusion

The overall results indicate that transformer-based models, particularly GPT-2, show promise in Chinese idiom completion tasks. The performance of the models was evaluated based on both fine-tuning and prompt engineering tasks, and it was observed that GPT-2 models outperformed all other models in terms of both accuracy and f1-score. The success of fine-tuning in improving the model's performance suggests that it is an effective method for enhancing model accuracy. Additionally, the findings suggest that using larger training sets can result in better model performance, indicating that it may be beneficial to use larger training sets in future work.

6.1 Limitations and Recommendations

The success of our project depends heavily on the choice of model. We found that the GPT-2 model performed the best in the idiom completion task and the prompt engineering task. However, there are many other models available, and it is possible that there are even more powerful models that we have not tested yet. For example, the GLM-130B model is an open bilingual pre-trained model that has shown good performance on the Chinese idiom comprehension and completion task (Zeng et al., 2022). Unfortunately, due to time constraints, we were not able to test this model. Therefore, further investigation into more powerful models can potentially improve the performance of our system.

We also observed a significant improvement in model performance when increasing the size of the training set. However, we were not able to investigate the limit of each model as the evaluation score can be largely affected by the size of the data. Future research should focus on exploring the effect of dataset size on model performance and determining the optimal size of the dataset for each model. Additionally, we recommend further investigation into the impact of other factors such as the choice of optimizer, learning rate, and batch size on the performance of our models. This can help us understand the limitations of our current approach and suggest possible improvements for future work.

6.2 Interpretations and Implications

In terms of fine-tuning, the performance of different models on the Chinese idiom completion task may be attributed to the differences in their architectures. BERT and T5 are both transformer-based models, but they may not be optimized for the specific task of Chinese idiom completion. Additionally, T5-small and mT5-small have smaller sizes and fewer parameters than their larger counterparts, which may limit their ability to capture the nuances of Chinese idioms. On the other hand, GPT-2 is also a transformer-based model, but it has a larger size and more parameters than the other models, which may allow it to capture more complex patterns and relationships in the data. Furthermore, GPT-2 uses a language modelling objective, which is similar to the Chinese idiom completion task and may have contributed to its better performance. Overall, the results suggest that larger transformer-based models may be more effective for Chinese idiom completion, and further research may be needed to optimize other models for this task.

The difference in performance between the models on prompt engineering can be attributed to their architecture and pre-training. T5 and its variants, which are designed for sequence-to-sequence tasks, did not perform well on this task, likely due to the fact that prompt engineering is not a natural fit for these models. Based on the 0 scores from T5-small and mT5-small, we can conclude that the proper prompts highly rely on the training data during fine-tuning. In other words, the models can only execute the same instructions as those in the training data. This shows that the instructions can be useless for both models, and we only need to provide the models with input and output results. On the other hand, GPT-2, which are based on the transformer architecture and pre-trained on large amounts of data, performed better on this task. This suggests that models with a strong pre-training regimen and more advanced architectures are better suited for prompt engineering tasks. Further, these results suggest that prompt engineering is not a one-size-fits-all solution and the effectiveness of this strategy may depend on the architecture and specific task requirements of the model. Overall, the results indicate that prompt engineering can be a useful technique for improving the performance of certain models on specific natural language processing tasks and that it is important to carefully consider the choice of model architecture when applying this technique.

References:

Chen, D., Bolton, J., Manning, C.D. (2016). A thorough examination of the CNN/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858

Cui, Y., et al. (2018). A span-extraction dataset for Chinese machine reading comprehension. arXiv preprint arXiv:1810.07366

Hermann, K.M., et al. (2015). Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems 28

Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E. (2017). Race: large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683

Luong, M.T., Pham, H., Manning, C.D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025

Sha, Y., Wu, M., Zeng, Z., Ge, X., Huang, Z., & Wang, H. (2023). A Prompt-Based Representation Individual Enhancement Method for Chinese Idiom Reading Comprehension. In *Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part III* (pp. 682-698). Cham: Springer Nature Switzerland.

Tan, M., Jiang, J. (2020). A BERT-based dual embedding model for Chinese idiom prediction. arXiv preprint arXiv:2011.02378

Wang, X., Zhao, H., Yang, T., Wang, H. (2020). Correcting the misuse: a method for the Chinese idiom cloze test. In: Proceedings of Deep Learning Inside Out (DeeLIO): the First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pp. 1–10.

Zeng, A. et al. (2022). GLM-130B:ANOPENBILINGUALPRE-TRAINED MODEL. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2210.02414.pdf

Zheng, C., Huang, M., & Sun, A. (2019). ChID: A large-scale Chinese IDiom dataset for cloze test. https://doi.org/10.48550/arxiv.1906.01265