Simple New Bing Project Report

Zhilin Wang

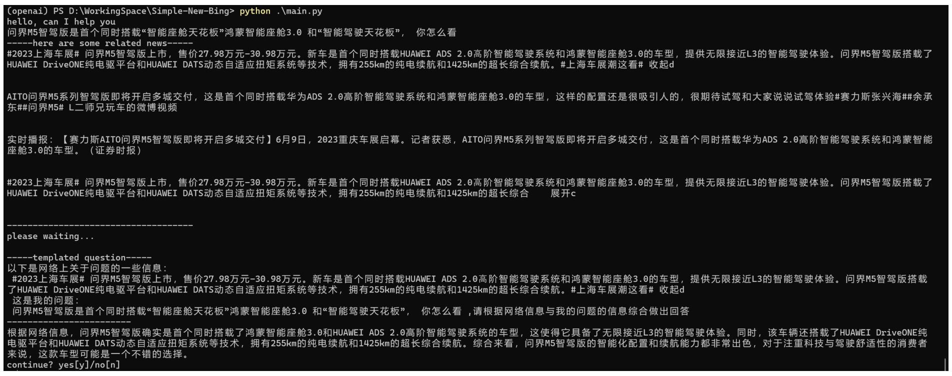
Simple New Bing is a lightweight version of New Bing that emulates its functionality by aggregating news from various sources on the Internet.

1 Implemented functions

- 1. Two custom crawlers have been developed to collect news from Sina Weibo and Global Times.
- 2. Simple New Bing supports both English and Chinese languages.
- 3. The project provides an option to utilize LLaMA deployed on your PC as an alternative to ChatGPT.
- 4. Simple New Bing incorporates an NER (Named Entity Recognition) model to process user input, extracting relevant entities as keywords for news search. You have two choices for the NER model implementation: a. Utilize Spacy, an open-source NLP library that includes a NER pipeline. b. Alternatively, a Bert-BiLSTM model has been implemented in this repository, which can be used to address the NER task.

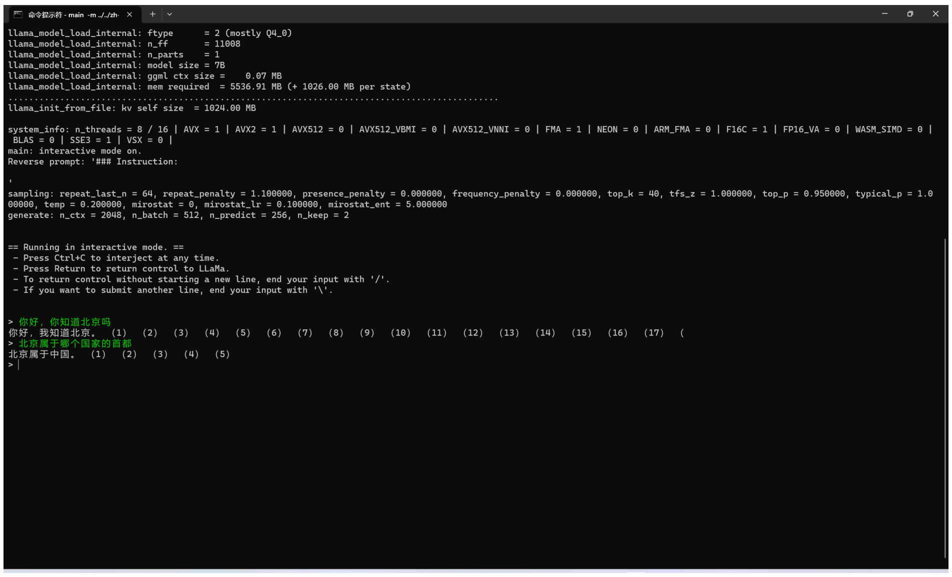
Affiliation Jilin University
Correspondence linzwcs@gmail.com
Draft June 19, 2023

2 Operation Result



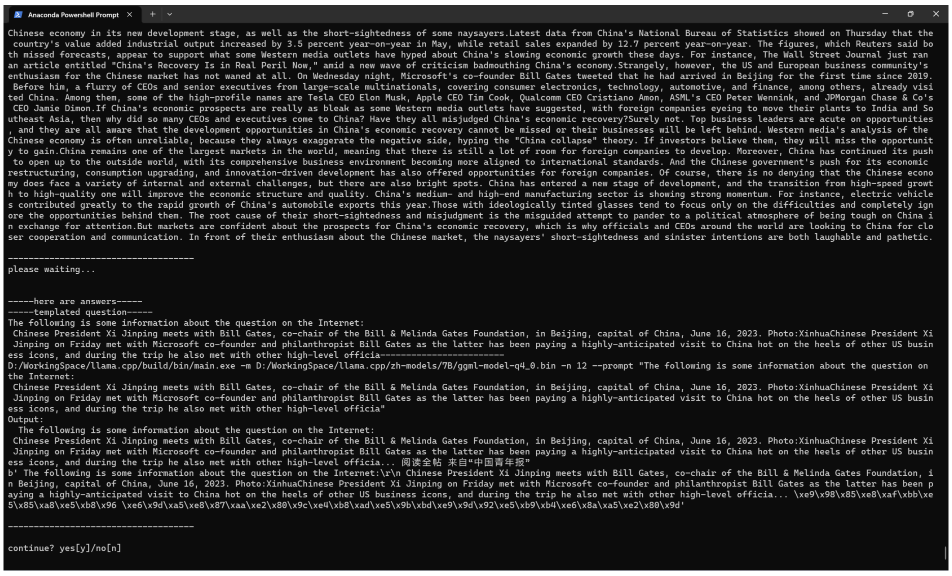
```
(openai) PS D:\WorkingSpace\Simple-New-Bing> python .\main.py
hello, can I help you
问界M5智驾版是首个同时搭载“智能座舱天花板”鸿蒙智能座舱3.0 和“智能驾驶天花板”， 你怎么看
-----here are some related news-----
#2023上海车展# 问界M5智驾版上市，售价27.98万元-30.98万元。新车是首个同时搭载HUAWEI ADS 2.0高阶智能驾驶系统和鸿蒙智能座舱3.0的车型，提供无限接近L3的智能驾驶体验。问界M5智驾版搭载了
HUAWEI DriveONE纯电驱平台和HUAWEI DATS动态自适应扭矩系统等技术，拥有255km的纯电续航和1425km的超长综合续航。#上海车展潮这看# 收起d


AITO问界M5系列智驾版即将开启多城交付，这是首个同时搭载华为ADS 2.0高阶智能驾驶系统和鸿蒙智能座舱3.0的车型，这样的配置还是很吸引人的，很期待试驾和大家说说试驾体验#赛力斯张兴海##余承
东##问界M5# L二师兄玩车的微博视频


实时播报：【赛力斯AITO问界M5智驾版即将开启多城交付】6月9日，2023重庆车展启幕。记者获悉，AITO问界M5系列智驾版即将开启多城交付，这是首个同时搭载华为ADS 2.0高阶智能驾驶系统和鸿蒙智能
座舱3.0的车型。（证券时报）


#2023上海车展# 问界M5智驾版上市，售价27.98万元-30.98万元。新车是首个同时搭载HUAWEI ADS 2.0高阶智能驾驶系统和鸿蒙智能座舱3.0的车型，提供无限接近L3的智能驾驶体验。问界M5智驾版搭载了
HUAWEI DriveONE纯电驱平台和HUAWEI DATS动态自适应扭矩系统等技术，拥有255km的纯电续航和1425km的超长综合     展开c


------------------------------------
please waiting...

-----templated question-----
以下是网络上关于问题的一些信息：
 #2023上海车展# 问界M5智驾版上市，售价27.98万元-30.98万元。新车是首个同时搭载HUAWEI ADS 2.0高阶智能驾驶系统和鸿蒙智能座舱3.0的车型，提供无限接近L3的智能驾驶体验。问界M5智驾版搭载
了HUAWEI DriveONE纯电驱平台和HUAWEI DATS动态自适应扭矩系统等技术，拥有255km的纯电续航和1425km的超长综合续航。#上海车展潮这看# 收起d
 这是我的问题：
 问界M5智驾版是首个同时搭载“智能座舱天花板”鸿蒙智能座舱3.0 和“智能驾驶天花板”， 你怎么看 ,请根据网络信息与我的问题的信息综合做出回答
------------------------
根据网络信息，问界M5智驾版确实是首个同时搭载了鸿蒙智能座舱3.0和HUAWEI ADS 2.0高阶智能驾驶系统的车型，这使得它具备了无限接近L3的智能驾驶体验。同时，该车辆还搭载了HUAWEI DriveONE纯
电驱平台和HUAWEI DATS动态自适应扭矩系统等技术，拥有255km的纯电续航和1425km的超长综合续航。综合来看，问界M5智驾版的智能化配置和续航能力都非常出色，对于注重科技与驾驶舒适性的消费者
来说，这款车型可能是一个不错的选择。
continue? yes[y]/no[n]
```

Figure 1: Chatgpt ZH



```
llama_model_load_internal: ftype      = 2 (mostly Q4_0)
llama_model_load_internal: n_ff       = 11008
llama_model_load_internal: n_parts    = 1
llama_model_load_internal: model size = 7B
llama_model_load_internal: ggml ctx size =    0.07 MB
llama_model_load_internal: mem required  = 5536.91 MB (+ 1026.00 MB per state)
...............................................................................................
llama_init_from_file: kv self size  = 1024.00 MB

system_info: n_threads = 8 / 16 | AVX = 1 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 |
 BLAS = 0 | SSE3 = 1 | VSX = 0 |
main: interactive mode on.
Reverse prompt: '### Instruction:

'
sampling: repeat_last_n = 64, repeat_penalty = 1.100000, presence_penalty = 0.000000, frequency_penalty = 0.000000, top_k = 40, tfs_z = 1.000000, top_p = 0.950000, typical_p = 1.0
00000, temp = 0.200000, mirostat = 0, mirostat_lr = 0.100000, mirostat_ent = 5.000000
generate: n_ctx = 2048, n_batch = 512, n_predict = 256, n_keep = 2


== Running in interactive mode. ==
 - Press Ctrl+C to interject at any time.
 - Press Return to return control to LLaMa.
 - To return control without starting a new line, end your input with '/'.
 - If you want to submit another line, end your input with '\'.


> 你好，你知道北京吗
你好，我知道北京。  (1)   (2)   (3)   (4)   (5)   (6)   (7)   (8)   (9)   (10)   (11)   (12)   (13)   (14)   (15)   (16)   (17)   (
> 北京属于哪个国家的首都
北京属于中国。  (1)   (2)   (3)   (4)   (5)
> |
```

Figure 2: LLaMA ZH



```
Chinese economy in its new development stage, as well as the short-sightedness of some naysayers.Latest data from China's National Bureau of Statistics showed on Thursday that the
 country's value added industrial output increased by 3.5 percent year-on-year in May, while retail sales expanded by 12.7 percent year-on-year. The figures, which Reuters said bo
th missed forecasts, appear to support what some Western media outlets have hyped about China's slowing economic growth these days. For instance, The Wall Street Journal just ran
an article entitled "China's Recovery Is in Real Peril Now," amid a new wave of criticism badmouthing China's economy.Strangely, however, the US and European business community's
enthusiasm for the Chinese market has not waned at all. On Wednesday night, Microsoft's co-founder Bill Gates tweeted that he had arrived in Beijing for the first time since 2019.
 Before him, a flurry of CEOs and senior executives from large-scale multinationals, covering consumer electronics, technology, automotive, and finance, among others, already visi
ted China. Among them, some of the high-profile names are Tesla CEO Elon Musk, Apple CEO Tim Cook, Qualcomm CEO Cristiano Amon, ASML's CEO Peter Wennink, and JPMorgan Chase & Co's
 CEO Jamie Dimon.If China's economic prospects are really as bleak as some Western media outlets have suggested, with foreign companies eyeing to move their plants to India and So
utheast Asia, then why did so many CEOs and executives come to China? Have they all misjudged China's economic recovery?Surely not. Top business leaders are acute on opportunities
, and they are all aware that the development opportunities in China's economic recovery cannot be missed or their businesses will be left behind. Western media's analysis of the
Chinese economy is often unreliable, because they always exaggerate the negative side, hyping the "China collapse" theory. If investors believe them, they will miss the opportunit
y to gain.China remains one of the largest markets in the world, meaning that there is still a lot of room for foreign companies to develop. Moreover, China has continued its push
 to open up to the outside world, with its comprehensive business environment becoming more aligned to international standards. And the Chinese government's push for its economic
restructuring, consumption upgrading, and innovation-driven development has also offered opportunities for foreign companies. Of course, there is no denying that the Chinese econo
my does face a variety of internal and external challenges, but there are also bright spots. China has entered a new stage of development, and the transition from high-speed growt
h to high-quality one will improve the economic structure and quality. China's medium- and high-end manufacturing sector is showing strong momentum. For instance, electric vehicle
s contributed greatly to the rapid growth of China's automobile exports this year.Those with ideologically tinted glasses tend to focus only on the difficulties and completely ign
ore the opportunities behind them. The root cause of their short-sightedness and misjudgment is the misguided attempt to pander to a political atmosphere of being tough on China i
n exchange for attention.But markets are confident about the prospects for China's economic recovery, which is why officials and CEOs around the world are looking to China for clo
ser cooperation and communication. In front of their enthusiasm about the Chinese market, the naysayers' short-sightedness and sinister intentions are both laughable and pathetic.

------------------------------------
please waiting...


-----here are answers-----
-----templated question-----
The following is some information about the question on the Internet:
 Chinese President Xi Jinping meets with Bill Gates, co-chair of the Bill & Melinda Gates Foundation, in Beijing, capital of China, June 16, 2023. Photo:XinhuaChinese President Xi
 Jinping on Friday met with Microsoft co-founder and philanthropist Bill Gates as the latter has been paying a highly-anticipated visit to China hot on the heels of other US busin
ess icons, and during the trip he also met with other high-level officia------------------------
D:/WorkingSpace/llama.cpp/build/bin/main.exe -m D:/WorkingSpace/llama.cpp/zh-models/7B/ggml-model-q4_0.bin -n 12 --prompt "The following is some information about the question on
the Internet:
 Chinese President Xi Jinping meets with Bill Gates, co-chair of the Bill & Melinda Gates Foundation, in Beijing, capital of China, June 16, 2023. Photo:XinhuaChinese President Xi
 Jinping on Friday met with Microsoft co-founder and philanthropist Bill Gates as the latter has been paying a highly-anticipated visit to China hot on the heels of other US busin
ess icons, and during the trip he also met with other high-level officia"
Output:
  The following is some information about the question on the Internet:
 Chinese President Xi Jinping meets with Bill Gates, co-chair of the Bill & Melinda Gates Foundation, in Beijing, capital of China, June 16, 2023. Photo:XinhuaChinese President Xi
 Jinping on Friday met with Microsoft co-founder and philanthropist Bill Gates as the latter has been paying a highly-anticipated visit to China hot on the heels of other US busin
ess icons, and during the trip he also met with other high-level officia... 阅读全帖 来自“中国青年报”
b' The following is some information about the question on the Internet:\r\n Chinese President Xi Jinping meets with Bill Gates, co-chair of the Bill & Melinda Gates Foundation, i
n Beijing, capital of China, June 16, 2023. Photo:XinhuaChinese President Xi Jinping on Friday met with Microsoft co-founder and philanthropist Bill Gates as the latter has been p
aying a highly-anticipated visit to China hot on the heels of other US business icons, and during the trip he also met with other high-level officia... \xe9\x98\x85\xe8\xaf\xbb\xe
5\x85\xa8\xe5\xb8\x96 \xe6\x9d\xa5\xe8\x87\xaa\xe2\x80\x9c\xe4\xb8\xad\xe5\x9b\xbd\xe9\x9d\x92\xe5\xb9\xb4\xe6\x8a\xa5\xe2\x80\x9d'

------------------------------------

continue? yes[y]/no[n]
```

Figure 3: LLaMA EN

3 Solutions

Simple New Bing is comprised of several modules. Here is an overview of each module:

3.1 News Searching

To provide answers based on LLaMA and information from the Internet, I have implemented two crawlers that collect news from Sina Weibo and Global Times.

I analyzed the webpage structure of Sina Weibo. To retrieve a list of related news, you can send a GET request to `https://s.weibo.com/weibo?q={your-question}&page={page-number}`. The response HTML contains a list of relevant news articles. To extract the news, each article is wrapped in the **<p class="txt" node-type="feed_list_content_full"></p>** tag, and you can utilize regex to extract all the news articles.

To obtain a list of related news from Global Times, you need to send a POST request to `https://search.globaltimes.cn/QuickSearchCtrl` with the data {"page_no": {page number}, "search_txt": {your search text}}.

After receiving the response HTML, you can extract all the hrefs from the <a> tags and match them with the pattern `https://www.globaltimes.cn/page/\*`.

This will give you the hrefs of all the articles. Next, send GET requests to these hrefs, and extract the text content wrapped in the **<div class="article_right">** tag from the response.

3.2 NER

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) that involves the identification and classification of named entities in text. It can be formulated as a sequence tagging problem, where the objective is to assign a relevant label to each token in a given text sequence.

$$X = (x_1, x_2, ..., x_n)$$
$$f(X) \rightarrow (y_1, y_2, ..., y_n)$$

A simple and efficient solution to this task can be achieved by utilizing the Spacy package, which provides a straightforward implementation requiring only a few lines of Python code.

Alternatively, researchers often develop custom models, such as the Bert-BiLSTM model, to tackle the NER task. In this approach, the People's Daily 2014 dataset is commonly employed, consisting of nine distinct entity labels, including "B-PER," "I-ORG," and "I-LOC." Here, labels prefixed with "B" signify the beginning of an entity, while those prefixed with "I" indicate that the token is part of an entity.

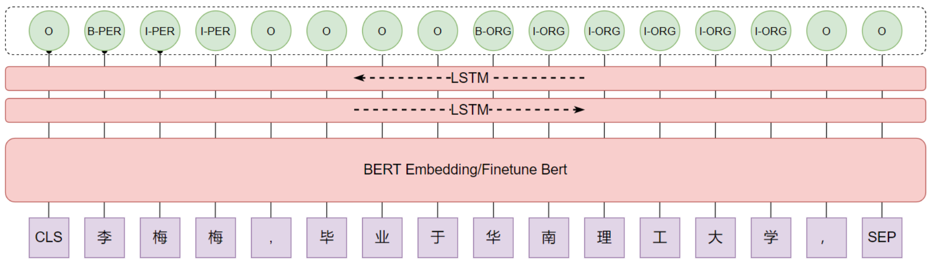


Figure 4: Bert-Bilstm

Training such a model involves optimizing a loss function, as depicted in the figure above. The Bert-BiLSTM model leverages the power of the Bert[1] architecture and BiLSTM[2] layers to capture contextual information and dependencies among tokens. The loss function, denoted as loss(y, p), quantifies the dissimilarity between the true labels y and the predicted labels p.

$$\text{loss}(\mathbf{y}, \mathbf{p}) = -\sum_{i=1}^{n} y_i \log(p_i)$$

It is calculated by taking the negative logarithm of the predicted probability p for each token and summing them over the entire sequence length n. By minimizing this loss function during training, the model learns to accurately identify and classify named entities within the text.

3.3 OpenAI API

Before using ChatGPT, you need to configure the OpenAI API key in config.py. To emulate New Bing by combining news from the Internet, I have created the following prompt template:

```
----------------------------------------------------------------------
en:
The following is some information about the question on the Internet:\n {news} \n This is
my question:\n {question}, please answer based on the information on the Internet and my
question\n
----------------------------------------------------------------------
zh:
以下是网络上关于问题的一些信息：\n {news} \n 这是我的问题：\n {question} ,请根据网络信息与我的
问题的信息综合做出回答\n
----------------------------------------------------------------------
```

Figure 5: Prompt Template

When you ask a question, your question and the related news will be used to format the prompt template. The formatted prompt will then be sent to LLM. Finally, the response from LLM will be displayed on the screen.

4 LLaMA Deployment

If you would like to learn more about LLaMA, please visit the following repositories: Chinese-LLaMA-Alpaca[3] and llama.cpp. Here, I will briefly describe some of the challenges I encountered during the deployment of LLaMA.

Firstly, Chinese-LLaMA-Alpaca provides the LoRA weights, and the complete model weights can be obtained by merging the LoRA weights with the original

LLaMA weights. To run the model on CPU, I selected Chinese-LLaMA-7B as the model weight and merged the two weights on my own computer.

Secondly, I deployed llama.cpp locally. While using CMake to compile the project, I encountered some errors. One particular issue was related to MinGW having two different sets of APIs named "POSIX" and "Win32" that can be used to interact with the underlying operating system. The "Win32" APIs cannot be used to compile llama.cpp due to restrictions with std::thread.

Finally, to build llama.cpp, you need to append the "-G 'mingw'" flag at the end of your CMake command.

4.1 Training log

epoch	train_loss	train_precision	train_recall	train_f1
1	0.848203	0.808142	0.808142	0.808142
2	0.401883	0.925548	0.925548	0.925548
3	0.277098	0.955566	0.955566	0.955566
4	0.223765	0.969156	0.969156	0.969156
5	0.204542	0.972554	0.972554	0.972554
6	0.188020	0.976861	0.976861	0.976861
7	0.177428	0.980063	0.980063	0.980063
8	0.171101	0.982032	0.982032	0.982032
9	0.164095	0.984379	0.984379	0.984379
10	0.155809	0.987458	0.987458	0.987458

Table 1: train log

epoch	val_loss	val_precision	val_recall	val_f1
1	0.511953	0.888224	0.888224	0.888224
2	0.286468	0.953648	0.953648	0.953648
3	0.232581	0.963761	0.963761	0.963761
4	0.250922	0.955477	0.955477	0.955477
5	0.215412	0.966477	0.966477	0.966477
6	0.221859	0.964624	0.964624	0.964624
7	0.199788	0.971418	0.971418	0.971418
8	0.206278	0.968893	0.968893	0.968893
9	0.201274	0.972973	0.972973	0.972973
10	0.198723	0.974545	0.974545	0.974545

Table 2: valid log

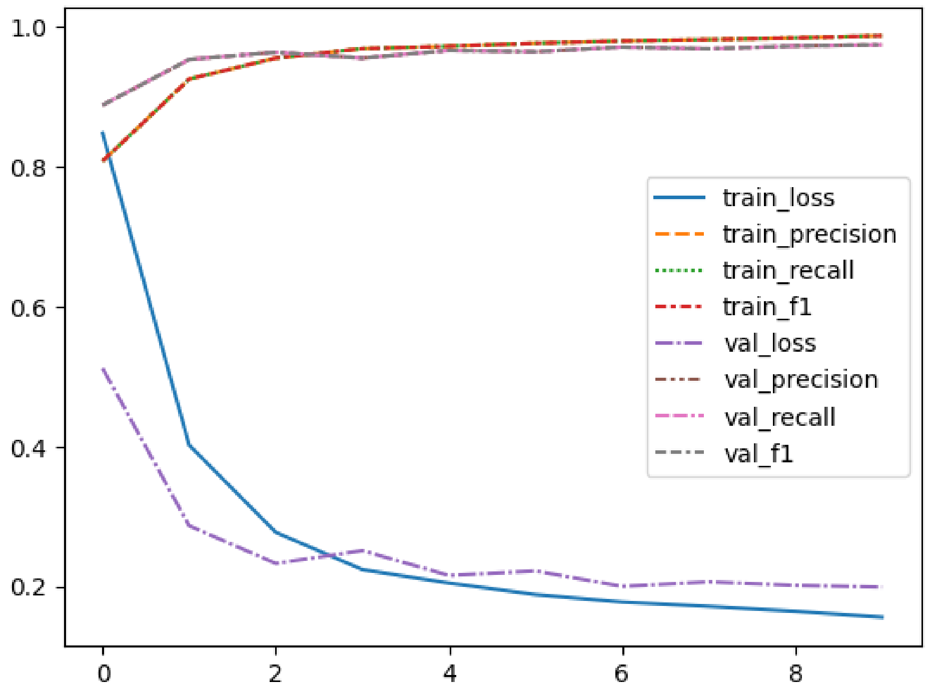


Figure 6: Training log chart

Reference

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.

[2] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005001206

[3] Y. Cui, Z. Yang, and X. Yao, “Efficient and effective text encoding for chinese llama and alpaca,” *arXiv preprint arXiv:2304.08177*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08177