Abstract

Nowadays, foundation models become one of fundamental infrastructures in artificial intelligence, paving ways to the general intelligence. However, the reality presents two urgent challenges: existing foundation models are dominated by the English-language community; users are often given limited resources and thus cannot always use foundation models. To support the development of the Chinese-language community, we introduce an open-source project, called Fengshenbang, which leads by the research center for Cognitive Computing and Natural Language (CCNL). Our project has comprehensive capabilities, including large pre-trained models, user-friendly APIs, benchmarks, datasets, and others. We wrap all these in three sub-projects: the Fengshenbang Model, the Fengshen Framework, and the Fengshen Benchmark.

An open-source roadmap, Fengshenbang, aims to re-evaluate the open-source community of Chinese pre-trained large-scale models, prompting the development of the entire Chinese large-scale model community. We also want to build a user-centered open-source ecosystem to allow individuals to access the desired models to match their computing resources. Furthermore, we invite companies, colleges, and research institutions to collaborate with us to build the large-scale open-source model-based ecosystem. We hope that this project will be the foundation of Chinese cognitive intelligence.

Note that this report also has a Chinese-language version (Starting from Section 6).

1 Introduction

Remarkable advances in Artificial Intelligence (AI) have produced great models, in particular, pre-training based foundation models (Bommasani et al., 2021) become an emerging paradigm. In contrast to traditional AI models that must be trained on vast datasets for one or a few scenarios, foundation models can be adapted to a wide range of downstream tasks, therefore, limiting the amount of resource demanded to acquire an AI venture off the ground. Moreover, we observe that these models grow rapidly within a short period, around 10 times each year. For instance, BERT (Devlin et al., 2019) has 100 million parameters and GTP-3 (Brown et al., 2020) has over 100 billion parameters. Many of the forefront challenges in AI, especially generalization ability, are becoming achievable due to this inspiring trend.

Foundation models, most notably language models, are dominated by the English-language community. The Chinese language as the world’s largest spoken language (native speakers), however, has no systematic research resources to support it, making the progress in the Chinese language domain lag behind others. To address this urgent need, we develop a Chinese language driven foundation ecosystem, named Fengshenbang, that incorporates pre-trained models, task-specific fine-tune applications, benchmarks, and datasets.

Our goal is to build a comprehensive, standardized and user-centered ecosystem. Although this can be instantiated in a variety of ways, we present the following design that we find to be particularly effective. Concretely, as illustrated in Figure 1, Fengshenbang ecosystem has three core modules as follows.

![Figure 1: The overview of Fengshenbang project.]
- Fengshenbang Model (Section 2)
- Fengshen Framework (Section 3)
- Fengshenbang Benchmark (Section 4)

Before discussing the Fengshenbang models, we first lay out some preliminaries. The NLP community has broad research interests, which can be categorized into two main types: general-purpose tasks and special-purpose tasks. The former includes Natural Language Understanding (NLU), Natural Language Generation (NLG), and Natural Language Transformation (NLT); the latter covers multimodality, specific domains, and others. Importantly, we argue that a platform with comprehensive foundation models is necessary, therefore, we consider all of these in Fengshenbang models. Further, we provide associated models to fine-tune for downstream tasks, so that users with limited resources can access foundation models without effort.

Next, we introduce the Fengshen framework to be user-centered, allowing users to further refine or modify models w.r.t their purposes based on our resources. More specifically, we provide a flexible menu in order to enable low-cost utilization, including standardized data processors, detailed tutorial examples, a docker-alike environment, and industry-standard APIs.

Last but not the least, our proposed ecosystem includes a benchmark module, allowing users to perform fair comparisons and the whole community to track progress. In particular, we choose to open source the leaderboard system in the future to make the comparison fair and promote the development of more customized Leaderboard systems.

To this end, we have introduced our ecosystem. Although this seems complicated, with only 3 sequential steps, users can build their applications based on our resources.

Step 1: Choosing a pre-trained Chinese NLP model from our open-source library of Fengshenbang Models.
Step 2: Employing Fengshen Framework to adjust the model by exploring the our tutorial examples.
Step 3: Evaluating on downstream tasks, such as Fengshenbang Benchmarks or custom tasks.

2 Fengshenbang Model

In this section, we will introduce the details of building Fengshenbang Models. As shown in Table 1, we present a User-Centered Taxonomy (UCT) (Section 2.1.1), aiming to classify the requirement of users and designing corresponding models. More details of these models are included in Section 2.1.2, which explains our model selection criteria to ensure the quality. In fact, we provide a nomenclature in Section 2.1.3 to reduce the misunderstanding as low as possible, where users can locate the desired models based on the nomenclature. More details of these models can be find from Section 2.2 to Section 2.7. Please refer to Appendix A for more details of 49 models.

2.1 Model Design

2.1.1 User-centered Taxonomy (UCT)

Advances in NLP have developed numerous powerful models from different perspectives, including research and applications. To understand the difference and track progress, there is an opportunity to standardize taxonomy in this field. However, models are often difficult to be categorized due to their complexity. For example, TinyBERT (Jiao et al., 2020) can be classified as “Encoder-only” in the model architecture, or it can be assigned as “Distillation” under model parameter reduction methods. To reduce misunderstanding, we introduce the User-centered Taxonomy (UCT), which consults numerous NLPers. In general demands, there are common NLP tasks, which are classified into Natural Language Understanding (NLU), Natural Language Generation (NLG), and Natural Language Transformation (NLT). Due to the fast development, NLP community brings special demands to the entire AI community, which are often assigned to MultiModal (MM), Domains and Exploration. More specifically, we assign a series name for each task. Note that we will update UCT timely according to the development of the NLP field.

Table 1: User-centered taxonomies with example models. “TBD” indicates “To Be Discussed”.

| Demand | Task | Series | Example Model |
|--------|------|--------|---------------|
| General | NLU  | Erlangshen | BERT, DeBERTa |
| NLG    | Wenzhong | Randeng   | GPT2, T5, BART |
| NLT    | Randeng | BioBERT   | TBD            |
| Special| MM   | Taiyi    | CLIP           |
| Domain | Yuyuan | Exploration | TBD            |
| Exploration | TBD | | TBD            |

Natural Language Understanding (NLU)

NLU tasks make use of syntactic and semantic analysis of text to understand the meaning of sentences. The syntax is related to the grammatical...
structure of a sentence, and semantics refers to its intended meaning. Relationships between words and phrases are also important as these will lead to different concepts. In addition, some problems in this task are difficult to solve even for humans. To evaluate the performance of NLU, several tasks are developed to ensure reliability:

- Semantic Matching
- Sentiment Analysis
- Natural Language Inference
- Entity Recognition
- Relationship Extraction
- Event Extraction
- Chinese Word Segmentation
- ...

**Natural Language Generation (NLG)**

Different from computer reading comprehension in NLU, NLG is concerned with developing computer systems that produce understandable writing. An NLG-capable system should be able to generate natural language by forming its ideas as opposed to transforming existing data. And, the generated text needs to be coherent and understandable to humans. We assign several NLG tasks as follows.

- Creative Writing
- Causal Reasoning
- Controlled Generation
- Multi-Step Reasoning
- ...

**Natural Language Transformation (NLT)**

We define NLT tasks as source-to-target transformation tasks. In contrast to NLG, NLT is based on the source objects and target objects. Language models require generating or transforming target objects by understanding source objects. Taking machine translation as an example, given a text in one language, an AI system needs to generate the corresponding text in another language. In summary, we list the NLT tasks as follows.

- Machine Translation
- Text Summarization
- Text Simplification
- Grammatical Error Correction
- Question Answering
- Dialogue System
- ...

**MultiModal (MM)**

Due to the growing demand for complex scenarios, unimodal models cannot handle multiple modalities. Transformer-based Pre-trained Language Models (PLMs) are widely used in fields like computer vision and audio processing due to their flexible architecture. In addition, cognitive intelligence requires intelligent systems to learn from multiple modalities, including text, images, and audio. To this end, several multimodal scenarios are introduced, such as text-to-image generation and multimodal semantic understanding.

- Text-to-image Generation
- Image Captioning
- Cross-modal Retrieval
- Visual Question Answering
- Automatic Speech Recognition
- Text-to-speech
- Voice Conversation
- Protein Structure Prediction
- ...

**Domain**

Notably, PLMs have achieved phenomenal success in a variety of specific domains. Continuous pre-training is a key advantage of constructing domain-specific models. Because the model is not trained from start, it results in less computational resource consumption. Some domains and several related models are listed below:

- Finance: FinBERT (Yang et al., 2020)
- Biomedical: BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2022b)
- Legal: LEGAL-BERT (Chalkidis et al., 2020), ALeaseBERT (Leivaditi et al., 2020)
- Programming: CoTexT (Phan et al., 2021), CodeBERT (Feng et al., 2020), GraphCodeBERT (Guo et al., 2021), CodeGPT-adapted (Lu et al., 2021), Codex (Chen et al., 2021)
- Academic: OAG-BERT (Liu et al., 2021), MathBERT (Peng et al., 2021), SciBERT (Beltagy et al., 2019)
- ...

**Exploration**

Together with other organizations, such as technology companies and universities, we will develop some experimental models in NLP.

**2.1.2 Model Selection**

Since many papers and models are proposed every year, we select only a few of them for pre-training and then open source, aiming to control the overall quality and use computing resources wisely. We select models based on the following rules:

- **Powerful.** Some models present astonishing performance on downstream tasks. These are ei-
ther published in English version or even not released. In addition, they are often not matching with for Chinese-language community, but need be extended or modified.

**Diversity.** Multiple models are widely used in various NLP tasks, such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and Transformer (Vaswani et al., 2017). Moreover, they are often to be extended and adapted easily. We include these models when considering diverse scenarios such as different downstream tasks, architectures, model sizes and pre-training methods.

**Usability.** Open-sourced models should easy to be understood and implemented in practice. Additionally, users can use some models out-of-the-box for desired downstream tasks.

### 2.1.3 Naming Convention

To better understand Fengshenbang models, we introduce the naming namespace with the following template:

\[
Name \in \{\text{Series} - \text{Model} - \text{Parameter} - \text{Extra}\}^N
\]

Here, *Series* is the name of series that come from a list of gods in a Chinese fiction called Fengshenbang (Hsun, 2000). Each series manages a category of NLP task. *Model* represents the structure of models, such as BERT and GPT-3. *Parameter* is the number of parameters and *N* is the total number of our open-sourced models. *Extra* indicates extra information of settings, such as fine-tuning on downstream datasets. For example, “Erlangshen-Roberta-110M-NLI” shows this is in Erlangshen series to solve NLU tasks. The model structure is based on RoBERTa with 110M parameters. After pre-training on Chinese datasets, we fine-tune it on NLI tasks.

### 2.2 Erlangshen (NLU)

The Erlangshen series is designed to solve NLU problems, including Megatron-BERT, ZEN, RoBERTa, DeBERTa, Longformer, UBERT and UnifiedMC.

#### 2.2.1 MegatronBERT

For training a billion-sized BERT, we follow the instructions of Megatron-LM (Shoeybi et al., 2019) and pre-train the BERT (Devlin et al., 2019) on WuDao Corpora (180 GB version) (Yuan et al., 2021). Given the Chinese grammatical structure and the difficulty of training in large-scale model, we apply the following four pre-training strategies to improve BERT.

1. **Whole Word Masking (WWM).** We adopt WWM (Cui et al., 2021) by considering the linguistic features of Chinese, which is to process whole Chinese words instead of individual Chinese characters in popular WordPiece tokenizer.

2. **Knowledge-based Dynamic Masking (KDM).** Instead of random masking in Masked Language Modeling (MLM), we attempt to mask token-rich semantic information to yield effective and efficient models.

3. **Sentence Order Prediction (SOP).** Referring to ALBERT (Lan et al., 2020), the LMs learn the inter-sentence information to get powerful representations. Hence, as a pre-training task, Erlangshen-MegatronBert models employ SOP instead of NSP.

4. **Pre-layer Normalization (Pre-LN).** After applying post-layer normalization in the LM pre-training phase, we find that the loss rises abnormally fast as the model size increases. Therefore, we apply Pre-LN (Xiong et al., 2020) to overcome this issue.

To the best of our knowledge, the largest BERT in the Chinese open source community is Erlangshen-MegatronBert-1.3B when publicly released. The Erlangshen-MegatronBert model has benefited many developers, as demonstrated by the 4.5K monthly downloads of our Erlangshen-MegatronBert model. Thanks to the excellent performance, Erlangshen-MegatronBert models gain three important mentions:

1. On November 10, 2021, it topped the FewCLUE few-shot learning tasks on CLUE benchmark (Xu et al., 2020). Among them, our model outperformed human performance in CHID (idiom fill-in-the-blank) and TNEWS (news classification) subtasks. In addition, it ranked the top in CHID (idiom fill-in-the-blank), CSLDCP (Subject Literature Classification), TNEWS (News Classification), IFLYTEK (Application Description Classification), CSL (Abstract Keyword Recognition), and CLUEWSC (Coreference Resolution) tasks.

2. On January 24, 2022, it topped the ZeroCLUE zero-shot Learning task on CLUE benchmark. For each of these tasks, we ranked the first places in CSLDCP (Subject Literature Classification), TNEWS (News Classification), IFLYTEK (Application Description Classification), CSL (Abstract Keyword Recognition), and CLUEWSC (Coreference Resolution) tasks.

3. We topped the CLUE benchmark semantic matching task on July 10, 2022, 2022 (Wang et al., 2022).

---

\(^1\)The data was obtained on August 18, 2022
Next, we will make the paper publicly available for details on Erlangshen-MegatronBert.

### 2.2.2 ZEN

We open source and publicly release ZEN1 (Diao et al., 2020) and ZEN2 (Song et al., 2021) using our Fengshen Framework in collaboration with the team ZEN. More precisely, by bringing together knowledge extracted by unsupervised learning, ZEN1 learns different textual granularity information through N-gram methods. ZEN1 can obtain good performance gains by training only on a single small corpus (low-resource scenarios). ZEN2 pre-trains the n-gram-enhanced encoders with large-scale datasets and special pre-training strategies. In the next step, we continue with the ZEN team to explore the optimization of PLM and improve the performance on downstream tasks.

- Erlangshen-ZEN1-224M-Chinese
- Erlangshen-ZEN2-345M-Chinese
- Erlangshen-ZEN2-668M-Chinese

### 2.2.3 RoBERTa

To obtain the Chinese version, we basically follow the instructions of RoBERTa (Liu et al., 2019). Considering the Chinese grammar, we adopt WWM in MLM and pre-train on the WuDao Corpora (180 GB version).

The RoBERTa has the following list:

- Erlangshen-Roberta-110M-NLI
- Erlangshen-Roberta-110M-Sentiment
- Erlangshen-Roberta-110M-Similarity
- Erlangshen-Roberta-330M-NLI
- Erlangshen-Roberta-330M-Sentiment
- Erlangshen-Roberta-330M-Similarity

### 2.2.4 DeBERTa

We mostly follow the instructions of DeBERTa-v2 (He et al., 2021) to obtain several Chinese versions. Considering the Chinese grammar, we adopt WWM in MLM and pre-train on WuDao Corpora (180 GB version) like Erlangshen-MegatronBert.

- Erlangshen-DeBERTa-v2-186M-Chinese-SentencePiece
- Erlangshen-DeBERTa-v2-320M-Chinese
- Erlangshen-DeBERTa-v2-710M-Chinese
- Erlangshen-DeBERTa-v2-97M-CWS-Chinese
- Erlangshen-DeBERTa-v2-97M-Chinese

### 2.2.5 Longformer

By following Longformer (Beltagy et al., 2020), we adopt WWM in MLM and pre-train on WuDao Corpora (180 GB version). Specifically, we adopt the Rotary Position Embedding (RoPE) (Su et al., 2021) to avoid the uneven sequence length of the pre-trained corpus.

- Erlangshen-Longformer-110M
- Erlangshen-Longformer-330M

### 2.2.6 UBERT

UBERT (Lu et al., 2022) was the winner solution in the 2022 AIWIN ARTIFICIAL INTELLIGENCE WORLD INNOVATIONS: Chinese Insurance Small Sample Multi-Task*. Our team, Fengshenbang, developed a unified framework based on BERT for multiple tasks and objectives. Our UBERT owns first place, as described in leaderboards (A and B). In addition to the unavailable datasets in the challenge, we carefully collect over 70 datasets from a variety of tasks for open-source UBERT. Besides out-of-the-box functionality, our UBERT can be employed in various scenarios such as NLI, entity recognition, and reading comprehension.

- Erlangshen-Ubert-110M-Chinese
- Erlangshen-Ubert-330M-Chinese

### 2.2.7 UnifiedMC (temporary)

We consider a new paradigm to employ the encoder-based pre-trained language models in zero-shot and few-shot scenarios. Our code and details about pre-training tasks will be made publicly available upon acceptance of the paper. In addition, our UnifiedMC models topped the FewCLUE and ZeroCLUE on August 30, 2022.

### 2.3 Wenzhong (NLG)

The Wenzhong series focus on solving NLG tasks.

#### 2.3.1 GPT2

We implement our GPT with 30 layers to obtain the powerful performance of Chinese GPT2 (Radford et al., 2019), which is larger than the original GPT2. Wenzhong-GPT2-3.5B is pre-trained on CLURCorpus2020 (Xu et al., 2020). The structure of Wenzhong2.0-GPT2-3.5B-chinese is the same.

*http://ailab.aiwin.org.cn/competitions/68
as Wenzhong-GPT2-3.5B, and is pre-trained on Wudao Corpus (300G version). Moreover, we implement a base size Wenzhong-GPT2-110M with 12 layers, which is pre-trained on Wudao Corpus (300G version).

- Wenzhong-GPT2-110M
- Wenzhong-GPT2-3.5B
- Wenzhong2.0-GPT2-3.5B-chinese

**2.4 Randeng (NLT)**

To handle NLT tasks, we introduce the Randeng series, which aims to solve source-to-target transformation tasks such as text summarization tasks.

**2.4.1 BART**

Based on BART (Lewis et al., 2020), we apply BERT tokenizer and WuDao Corpora (180 GB version) to train a Chinese version. Since the BERT tokenizer usually performs better than others for Chinese tasks, we employ it in one of Randeng-BART models.

- Randeng-BART-139M
- Randeng-BART-139M-SUMMARY
- Randeng-BART-759M-Chinese-BertTokenizer

**2.4.2 MegatronT5**

To get a large-scale T5 (Raffel et al., 2020), we make use of Megatron-LM (Shoeybi et al., 2019) method and WuDao Corpora (180 GB version) for pre-training.

- Randeng-MegatronT5-770M

**2.4.3 PEGASUS**

To solve Chinese text summarization tasks, we follow the PEGASUS (Zhang et al., 2020) guidelines. We employ the WuDao Corpora (180 GB version) as a pre-training dataset. In addition, considering that the Chinese sentence piece tokenization is unstable, we utilize jieba [3] and BERT tokenizer in our Randeng-PEGASUS.

- Randeng-Pegasus-238M-Chinese
- Randeng-Pegasus-238M-Summary-Chinese
- Randeng-Pegasus-523M-Chinese
- Randeng-Pegasus-523M-Summary-Chinese

**2.4.4 mT5**

We implement mT5 (Xue et al., 2021) for Chinese. In order to accelerate training, we only retrain the vocabulary and embedding corresponding to Chinese and English, and Corpus-Adaptive Pre-Training (CAPT) on the WuDao Corpora (180 GB version).

- Randeng-T5-77M
- Randeng-T5-784M

**2.4.5 Transformer-Denoise**

We explore a Chinese Transformer model to solve denoise tasks. We first pre-trained Transformer-XL on the Wudo corpus (180G version), and then fine-tuned it on a denoised dataset (developed by us). The denoise task is to reconstruct a fluent and clean text from a noisy input which includes random insertion/swap/deletion/replacement/sentence reordering.

- Randeng-Transformer-1.1B-Denoise

**2.5 Taiyi (MM)**

To handle multiple modalities, the Taiyi series is introduced and applied in cross-modal scenarios. In detail, we are working on protein structure prediction, speech-text representations and so on.

**2.5.1 CLIP**

We follow the experimental setup of CLIP (Radford et al., 2021) to obtain powerful visual-language intelligence. To obtain the CLIP for Chinese, we employ chinese-roberta-wwm (Cui et al., 2021) for the language encoder, and apply the ViT in CLIP for the vision encoder. We freeze the vision encoder and tune the language encoder to speed up and stabilize the pre-training process. Moreover, we apply Noah-Wukong dataset (Gu et al., 2022a) and Zero-Corpus (Xie et al., 2022) as the pre-training datasets. To the best of our knowledge, our Taiyi-CLIP is currently the only open-sourced Chinese CLIP in the huggingface community.

Moreover, we apply the Taiyi-CLIP models in various text-to-image generation scenarios, which shows the powerful capability of Chinese language understanding, such as generating images with Chinese ancient poems.

- Taiyi-CLIP-RoBERTa-326M-ViT-H-Chinese
- Taiyi-CLIP-Roberta-102M-Chinese
- Taiyi-CLIP-Roberta-large-326M-Chinese

**2.5.2 MAP (temporary)**

We propose an exploratory multimodal model to obtain powerful embeddings. We design a special module to account for multimodal uncertainty and then pre-train the model with special pre-training

[3]https://github.com/fxsjy/jieba
strategies. The fine-tuned models are applied to challenging downstream tasks and achieve state-of-the-art performance. Our framework also helps uni-modal downstream tasks. Our code and details about pre-training tasks will be made publicly available upon acceptance of the paper.

- Taiyi-Roberta-124M-D
- Taiyi-Roberta-124M-D-v2
- Taiyi-vit-87M-D

2.6 Yuyuan (Domain)

Despite being open-sourced, several models still have some difficult issues to solve. It is necessary to improve the quality and quantity of existing open source domain-specific datasets. But domain-specific model design differs from general PLMs, for example processing proper names. For example, for existing models in finance such as FinBERT (Yang et al., 2020), the vocabularies of it and general BERT (Devlin et al., 2019) share 41% common tokens. Because of the lack of domain-specific terms in the vocabulary, many domain-specific words are incorrectly represented, damaging the learning of the model. Therefore, it is crucial to provide high-quality domain-specific PLMs.

2.6.1 BioBART

For the biomedical domain, we adopt BART as presented in our BioBART paper (Yuan et al., 2022). With Doamin-Adaptive Pre-Training (DAPT), we employ BART-large on PubMed abstracts⁴, which contains around 41 GB of biomedical research paper abstracts. Next, we collect various biomedical language generation tasks including dialogue, summarization, entity linking, and named entity recognition. We demonstrate that BioBART presents large improvements on different benchmarks and achieves competitive or superior results over existing state-of-the-art methods.

- Yuyuan-Bart-139M
- Yuyuan-Bart-400M

2.6.2 GPT2

We adopt the same architecture as Wenzhong-GPT2-3.5B to be pre-trained on 50 GB medical (PubMed) corpus. Our Yuyuan-GPT2-3.5B is the largest open-source GPT2 model in the medical domain. We further allow the model to judge facts by computing perplexity (PPL). To accomplish question-and-answer functionality, we transform the phrase pattern from interrogative to declarative.

- Yuyuan-GPT2-3.5B
- YuyuanQA-GPT2-3.5B

2.7 TBD (Exploration)

We will have some experimental explorations with several organizations. For example, the following models are the outcome of our joint efforts with Zhuiyi Technology.

- Zhouwenwang-Unified-1.3B
- Zhouwenwang-Unified-110M

3 Fengshen Framework

To address the issues in Section 1, we integrate the advantages of Huggingface (Wolf et al., 2019), Megatron-LM (Shoeybi et al., 2019), PyTorch-Lightning, and DeepSpeed (Rajbhandari et al., 2020) into Fengshen framework. If the users are familiar with the above frameworks, then they can use our deep learning framework without any effort. Our framework can be utilized to pre-train large-scale models even over 10 billion parameters by using terabytes level data and be fine-tuned a range of downstream tasks. With this configuration, users can simply undertake distributed training and memory saving approaches, allowing them to focus more on model implementation and innovation. The Fengshen framework also provides codes and examples for open-source models in Huggingface and their applications.

Our framework has the following advantages:

- Superior performance compared to the original Torch, such as 300% acceleration in training.
- Support large-scale models: over 10 billion-level models in training and fine-tuning.
- Large datasets (Terabyte-level) are supported.
- The training process is easy to use by providing rich pre-training and downstream examples.
- Adapt to varied device environments (various devices such as CPU, GPU, and TPU).
- Support distributed training methods such as DDP and Zero Optimizer without any code modifications.

3.1 Architecture

Core components in the Fengshen Framework are presented in Figure 2: The core functionality of the model is to adapt to each module. It is straight-
forward to follow 3 steps to apply our Fengshen Framework:

1. Encapsulated Data Processing Flow
2. Encapsulate the model structure
3. Configure several plugins

3.2 Document

This paper provides the usage details\(^5\) of distributed model training, fine-tuning, and various large scale model applications. Then, we present the paper and various real-world code and tutorials from our participation competition.

3.3 Case Study

As shown in Fig. 3, only a few lines are required to construct the models in the Fengshen framework. For users who are already using frameworks like Huggingface, there is almost no cost to get started.

4 Fengshenbang Benchmark

Benchmarks are necessary to evaluate the capability of the models. It should be noted that Chinese is linguistically distinct from English and other Indo-European languages. Thus, we introduce Fengshenbang Benchmarks, which are composed of Chinese leaderboard system with different types of natural language tasks. Besides considering fair environments in several tasks, we plan to release a general pre-training dataset.

4.1 Task Criteria

To collect high-quality and robust benchmarks, we consider different aspects of testing the models. As a result, we identify the following requirements while building the Fengshenbang benchmark:

**Widely evaluated.** While some existing datasets are not designed in Chinese, they have been used extensively in NLP for years, e.g. SuperGLUE (Wang et al., 2019). We will gather some professional English and Chinese linguists to meticulously translate these popular datasets.

**Future-oriented.** In fact, a few NLP models already surpass human performance on several benchmarks. This declares that AI has reached or even can surpass human cognitive intelligence. One reason we believe is their limited scope of evaluation. A more urgent and necessary work is to construct challenging datasets instead of fitting existing datasets to 100% accuracy. Future benchmarks need to consider broader ethical, technical, and societal challenges. Our datasets will be published soon to better support the research community.

**Applicable.** Benchmarks are required to represent real-world scenarios. This allows us to collaborate with industry-active companies to publish datasets and collect real-world data.

4.2 Leaderboard

4.2.1 Chinese-SuperGLUE

As a powerful benchmark, the SuperGLUE is widely adopted, but it is only available in English. These high-quality datasets are hoped for by the Chinese community as well. The Chinese-language-oriented model can be challenged due to the lack of validation on the English dataset. This is very unfair and may result in the research being limited to a monolingual setting. One reason is that it is difficult to evaluate the quality of different datasets, leading reviewers to be more surprised about some unseen datasets. To address the above issues, we develop Chinese-SuperGLUE to evaluate Chinese models. The related paper and the leaderboard are coming soon.

4.2.2 QAKM

To evaluate the knowledge level in NLP models, we propose QAKM (Question Answering with Knowledge Models). The knowledge models are required
to learn domain-specific knowledge and answer unseen questions given a dataset. The task has been included in the NLPCC\(^6\), which is accessible in this website\(^7\).

5 Summary and Future Work

This report introduces our open source project Fengshenbang, aiming to build the foundation of Chinese Cognitive Intelligence. Our three sub-projects (Fengshenbang Model, Fengshen Framework, Fengshenbang Benchmark) support different aspects of intelligence systems’ progress. In addition, we want to emphasize that our project Fengshenbang is ongoing, i.e., we keep all sub-projects updated forward. When building a community, we expect individual and organizational contributors to join and refine the project together. The world needs a few good ideas.

Ethical Considerations

However, since our Fengshenbang Project provides the entire ecosystem to use, produce, and evaluate large-scale PLMs, we note that many companies and research institutions have deployed these models. Our models and benchmarks transfer gradually to the real world, thus having unpredictable impact on humans. There are many aspects to be concerned with considering the ethical impact: implicit bias of large-scale models, potential environmental issues, undesirable prejudgement in labeled data, inappropriate use of open source frameworks, etc. For a better understanding and deeper discussion of ethical issues, we will provide a detailed study in the next version of the report. While we encourage any developer to use anything from the Fengshenbang project to open debate on its utilization, such as task selection and deployment, we hope that this would reduce the chance of any misconduct.

Acknowledgements

We thank the CCNL GTS team in IDEA for providing feedback and help.

---

\(^6\)http://tcci.ccf.org.cn/conference/2022/cfpt.php
\(^7\)https://idea.edu.cn/ccnl-act/nlpcc-track1.html
引言

人工智能 (Artificial Intelligence, AI) 的显著进步产生了许多伟大的模型，特别是基于预训练的基础模型 (Bommasani et al., 2021) 成为了一个新兴的范式。传统的 AI 模型必须要在专门的巨大的数据集上一个或几个有限的场景进行训练。相比之下，基础模型可以适应广泛的下游任务。因此，基础模型让低资源的场景有 AI 落地的可能。并且，我们观察到这些模型的参数量正在以每年 10 倍的速度增长。比如，于 2018 年，BERT (Devlin et al., 2019) 的参数量仅有 1 亿量级。但是到了 2020 年，GPT-3 (Brown et al., 2020) 的参数量就已达到百亿的量级。由于这一鼓舞人心的趋势，人工智能中的许多前沿挑战，尤其是强大的泛化能力，正在变得可以被实现。

基础模型，尤其是语言模型，如今正在被英文社区主导。然而，中文作为这个世界上最大的口语音语种（母语者中），却缺乏系统性的研究资源支撑，这使得中文领域的研究进展相较于英文来说有些滞后。为了了解中文领域研究进展滞后和研究资源严重不足的问题，我们提出了一些名为“封神榜”的中文驱动的基础生态系统，其中包括了预训练模型、特定任务的微调应用，基准和数据集等。

我们的目标是构建一个全面的、标准化的，以用户为中心的生态系统。尽管这一目标可以通过多种方式去实现，但是我们经过对中文社区的重新审视与思考，提出了我们认为最为有效的方案。具体地，沈向洋在 IDEA 大会上正式宣布启动“封神榜”项目，旨在成为中文认知智能的基础设施。并且，我们希望通过全面开放和协作的方式共建开源社区，并推动中文自然语言技术发展。如图 4 所示，封神榜包括了三个核心模块：

- 封神榜模型(第 7 节)
- 封神框架(第 8 节)
- 封神榜单(第 9 节)

在具体讨论封神榜模型之前，我们先引入一些必要的概念。NLP 社区有着广泛的研究任务，这些任务可以被分为两类：通用任务和特殊任务。前者包括了自然语言理解（NLU），自然语言生成（NLG）和自然语言转换（NLT）任务。后者涵盖了多模态、特定领域等任务。重要的是，我们认为一个综合的基础模型平台是必要的。因此，我们在封神榜中考虑了上述所有任务。此外，我们还提供了在下游任务上微调好的相关模型。这使得资源有限的用户也可以轻松使用我们的基础模型。

接着，我们引入以用户为中心的封神框架。用户可以根据提供的资源进一步完善甚至修改模型。具体地，我们提供了简单灵活的使用方案，用户可以低成本地使用：包含标准化的数据库、详细的教程示例、类 docker 的环境和行业标准的 API 等。

最后，我们提出的生态系统中还包括了一个基准模块，它允许用户进行公平的比较并且可以给该系统追踪最新进展。特别地，我们在未来会发布榜单管理系统，希望可以推动更多定制的排行榜系统的发展。

以上三个模块就是我们的整个生态系统了。尽管这看起来可能有些复杂，但是只需三步，用户就可以根据我们的资源轻松构建所需的模型。

步骤 1：从我们的封神榜模型库中选择一个训练好的中文 NLP 模型。
步骤 2：通过阅读我们的教程示例，使用封神框架调整模型。
步骤 3：评估下游任务，如封神榜单或者自定义任务。

封神榜模型

在这一节中，我们将介绍封神榜模型相关的细节。如表 2 所示，我们展示了一个以用户为中心的分类学 (User-Centered Taxonomy, UCT)。UCT 会专门对用户的需求进行分类，提供相对应的模板和相关模型的设计指导。这些模型更多的细节将会在第 7.1.2 节中说明。用户可以根据我们的模型选择标准快速的定位到对应自身需求的模型。事实上，我们还提供了第 7.1.3 节中所述的命名法，可尽可能地降低寻找所需模型的难度。具体的设计说明以及对应的模型可以在第 7.2 节到第 7.7 节中找到。此外，我们所有的 49 个模型都可以在附录 A 中查到。

7.1 模型设计

7.1.1 以用户为中心的分类学 (UCT)

NLP 的先进研究中已经出现了大量从不同角度（研究和应用等）的模型。为了理解模型的不同以及跟踪社区的进展，我们希望可以
建立一个应用于NLP领域的标准的分类法。然而，因为模型通常由于其复杂性而难以分类。比如，TinyBERT (Jiao et al., 2020) 在模型构建中可以被归为“Encoder-only”，也可以在模型参数缩减方法中归为“Distillation”。为了减少误解，我们引入了以用户为中心的分类学(UCT)，它来自于众多NLPers的意见和咨询。现有的需求通常可以分为通用需求和特殊需求。在通用需求中，有常见的NLP任务，其中分为自然语言理解(NLU)、自然语言生成(NLG)和自然语言转换(NLT)任务。由于快速的发展，NLP社区给整个AI社区带来了特殊的需求，这些需求涵盖了多模态、特定领域和探索等。这个分类并不会止步不前，而是会随着用户需求的变化而与时俱进。

自然语言理解(NLU)
NLU任务要求模型利用文本的句法和语义进行分析来理解句子的含义。句法和句子的语法结构相关，语义则代表了其预期含义。单词和短语之间的关系也很重要，因为它们会导致概念理解上的差异。此外，这项任务中的一些问题即使是人类也难以解决。为了评估模型的NLU能力，如下的下游任务被设计出来以检验模型的可靠性。
- 语义匹配
- 情感分析
- 自然语言推理
- 实体识别
- 关系抽取
- 事件抽取
- 中文分词
- …

自然语言生成(NLG)
与NLU中的机器阅读理解能力不同，NLG专注于开发能够创作可被人理解的写作内容的计算系统。具有NLG能力的系统应该能够通过自身或所学的思想来生成自然语言，而不是转换现有的数据形式。而且，生成的文本需要连贯且易于理解。一些NLG任务如下:
- 创意写作
- 因果推理
- 可控生成
- 多步推理
- …

自然语言转换(NLT)
NLT任务是我们定义的一种全新的任务类型，是一种源→目标的转换任务。不同于NLG基于自己思想的生成，NLT要求语言模型在理解源文本对象的基础上，生成或者转换出目标文本对象。以机器翻译为例，给定一种语言的文本，AI系统需要生成另一种语言的相应文本。总之，我们将部分NLT任务在下面列出。
- 机器翻译
- 文本摘要
- 文本简化
- 语法纠错
- 问答系统
- 对话系统
- …

多模态(MM)
如今，复杂场景的需求在不断增长，而单模态的模型无法处理这些场景中的多模态信息。基于Transformer的预训练语言模型(Pre-trained Language Models, PLMs)因其灵活的架构而广泛用于计算机视觉和音频处理等领域。此外，认知智能需要智能系统从多种模式中学习，包括文本、图像和音频。为此，我们引入了多个多模态场景，如文本生成图像和多模态语义理解等。一些多模态的任务如下:
- 文本生成图像
- 图像标注
- 跨模态检索
- 视觉问答
- 自动语音识别
- 文字转语音
- 语音会话
- 蛋白质结构预测
- …

领域模型(Domain)
值得注意的是，PLMs在各个特定领域中都取得了惊人的成功。持续的预训练是特定领域模型的关键优势，因为模型不是从头开始预训练的，所以计算资源消耗更少。下面列出了一些领域和几个相关模型:
- 金融: FinBERT (Yang et al., 2020)
- 生物医学: BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2022b)
- 法律: LEGAL-BERT (Chalkidis et al., 2020), ALeaseBERT (Leivaditi et al., 2020)
• 编程：CoTeXT (Phan et al., 2021), CodeBERT (Feng et al., 2020), GraphCodeBERT (Guo et al., 2021), CodeGPT-adapted (Lu et al., 2021), Codex (Chen et al., 2021)
• 学术：OAG-BERT (Liu et al., 2021), MathBERT (Peng et al., 2021), SciBERT (Beltagy et al., 2019)
• …

探索
我们也希望和其他组织，如技术公司和大学，一起开发一些关于NLP的实验性模型。

7.1.2 模型选择
由于每年都有许多论文和模型被提出，为了控制整体质量和合理使用计算资源，我们只能选择其中的一部分进行预训练，然后进行开源。
所以，我们根据以下规则来选择模型：
强大的。一些模型在下游任务上呈现出惊人的性能，但是，这些模型要么是英文发表，要么甚至没有发布出来。此外，它们往往和中文社区不匹配，但需要扩展或修改。
多样性。 有很多模型已经被广泛地应用于多任务的NLP任务中了，比如BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020)和Transformer (Vaswani et al., 2017)。
并且，他们通常易于拓展和调整。我们采纳这些模型的时候，会考虑到下游任务的架构、模型大小和训练方法等不同情况。
实用的。 开放源码的模型应该容易被理解并且是可以实现的。此外，用户可以开箱即用地使用模型于所需的下游任务。

7.1.3 命名规则
为了更好地理解封神榜模型，我们引入了一个新的命名空间，模板如下：

\[
Name \in \{\text{Series–Model–Parameter–Extra}\}^N \quad (2)
\]

其中，Series 是取自中国小说《封神榜》中神话人物的名字 (Hsun, 2000)。每个系列对应着一类NLP任务。Model 是模型的结构，例如BERT 和 GPT-2。Parameter 是参数的数量，N 是我们开源模型的总数。Extra 是额外的信息，例如，在特定下游任务的数据集上微调过。代表了二郎神，给出一个例子，“Erlangshen-SciBERT-1.3M-999”代表了这款二郎神模型，用于解决自然语言理解(NLU)任务的。模型结构基于具有110M参数的RoBERTa模型。在与中文数据集进行预训练后，我们继续在自然语言推理(NLI)任务上对其进行微调。

7.2 二郎神(NLU)
二郎神系列旨在解决NLU问题，目前包括Megatron-BERT, ZEN, RoBERTa, DeBERTa, Longformer, UBERT和UnifiedMC等模型。

7.2.1 MegatronBERT
为了训练十亿级规模的BERT，我们参考Megatron-LM (Shoeybi et al., 2019)的方法，在悟道数据库(180 GB 版本) (Yuan et al., 2021)上预训练BERT (Devlin et al., 2019)。鉴于中文语法结构和大规模模型训练的难度，我们应用以下四种预训练策略来改进BERT (Devlin et al., 2019)。

(1) 整词掩码(Whole Word Masking, WWM)。
考虑到中文的语言特性，我们采用WWM (Cui et al., 2021)，即在WordPiece分词器中处理整个中文单词而不是单个汉字。
(2) 基于知识的动态掩码(Knowledge-based Dynamic Masking, KDM)。替换原本的掩码语言模型(MLM)中的随机掩码，我们尝试去遮掩那些具有更丰富语义的信息以产生有效和高效的模型。
(3) 句序预测(Sentence Order Prediction, SOP)。
参考ALBERT (Lan et al., 2020)，语言模型可以学习句间信息以获得强大的表征。因此，Erlangshen-MegatronBert模型使用SOP而不是NSP。
(4) 前层归一化(Pre-LN)。在语言模型预训练阶段使用层后归一化时，我们发现随着模型大小的增加，损失会异常上升的问题。因此，我们应用Pre-LN (Xiong et al., 2020)来克服这个问题。

据我们所知，在开源Erlangshen-MegatronBert-1.3B的时候，它是当时中国开源社区中参数量最大的BERT模型。Erlangshen-MegatronBert在Hugging Face上有着每月4.5K8的下载量，帮助了许多的开发人员。

因为其出色的表现，Erlangshen-MegatronBert获得了三个重要的成就：
(1) 2021年11月10日，它在CLUE基准(Xu et al., 2020)的FewCLUE上取得第一。其中，它在CHIDF(成语填空)和TNEWS(新闻分类)子任务中的表现优于人类表现。此外，它在CHIDF(成语填空), CSLDPC(学科文献分类), OCNLI(自然语言推理)任务中均名列前茅。
(2) 2022年1月24日，它在CLUE基准测试中的ZeroCLUE中取得第一。对于这些任务中的每一个，我们在CSLDCP(主题文献分类), TNEWS(新闻分类), IFLYTEK(应用描述分类), CSL(抽象关键字识别)和CLUEWSC(指代消
7.2.2 ZEN

We work with ZEN’s team to release ZEN1 (Diao et al., 2020) and ZEN2 (Song et al., 2021) models. These models can be built by using a similar method to obtain different text pieces. ZEN1 can be used in more than one language environment; ZEN2 uses a large-scale dataset and specific pretraining strategies to improve the performance of the models. Furthermore, we will continue to explore the ZEN team’s PLM optimization and help downstream tasks.

- Erlangshen-ZEN1-224M-Chinese
- Erlangshen-ZEN2-345M-Chinese
- Erlangshen-ZEN2-668M-Chinese

7.2.3 RoBERTa

To release an official version of PLMs, we follow Li et al.’s (2019) method. We use the RoBERTa model to train different tasks in the MLM setting. We also use WWM and other large datasets to improve performance.

- Erlangshen-ROBERTA-110M-Chinese
- Erlangshen-ROBERTA-330M-Chinese

7.2.4 DeBERTa

We follow He et al.’s (2021) method to release DeBERTa-v2. We use the DeBERTa model to train different tasks in the MLM setting. We also use WWM and other large datasets to improve performance.

- Erlangshen-DeBERTa-v2-186M-Chinese-SentencePiece
- Erlangshen-DeBERTa-v2-320M-Chinese
- Erlangshen-DeBERTa-v2-710M-Chinese
- Erlangshen-DeBERTa-v2-97M-CWS-Chinese
- Erlangshen-DeBERTa-v2-97M-Chinese

7.2.5 Longformer

Following Longformer’s (Beltagy et al., 2020) design, we release Longformer models in the MLM setting and use WWM for training. We also use other large datasets to improve performance.

- Erlangshen-Longformer-110M
- Erlangshen-Longformer-330M

7.2.6 UBERT

UBERT (Lu et al., 2022) is a new model released by AIWIN this year. It is designed for large-scale tasks and can achieve state-of-the-art performance.

- Erlangshen-UBERT-110M-Chinese
- Erlangshen-UBERT-330M-Chinese

7.2.7 UnifiedMC (Temporary)

We propose a new model, UnifiedMC, which can be used in different scenarios and achieve state-of-the-art performance.

- Erlangshen-UnifiedMC-110M-Chinese
- Erlangshen-UnifiedMC-330M-Chinese

7.3 NLG

We focus on improving NLG tasks.

7.3.1 GPT2

We achieve state-of-the-art performance on GPT2 in the CLUER corpus.

- Wenzhong-GPT2-3.5B-chinese
- Wenzhong-GPT2-3.5B
- Wenzhong-GPT2-3.5B

---

http://ailab.aiwin.org.cn/competitions/68
7.4 燃灯(NLT)
为了处理NLT任务，我们引入了燃灯系列，即目标是解决源对象到目标对象的转换任务，比如文本摘要任务。

7.4.1 BART
基于BART (Lewis et al., 2020)，我们应用BERT的分词器和悟道语料库(180G版本)训练了一个中文版本。因为BERT分词器通常在中文任务中表现比其他分词器好，所以我们在一个Randeng-BART模型中使用了它。
- Randeng-BART-139M
- Randeng-BART-139M-SUMMARY
- Randeng-BART-759M-Chinese-BertTokenizer

7.4.2 MegatronT5
为了得到一个大规模的T5 (Raffel et al., 2020)，我们使用了Megatron-LM (Shoeybi et al., 2019)的方法和悟道语料库(180G版本)用于预训练。
- Randeng-MegatronT5-770M

7.4.3 PEGASUS
为了解决中文的自动摘要任务，我们遵循PEGASUS (Zhang et al., 2020)的设计来训练中文的版本。我们使用了悟道语料库(180G版本)作为预训练数据集。此外，考虑到中文sentence piece不稳定，我们在Randeng-PEGASUS中同时使用了结巴分词和BERT分词器。
- Randeng-Pegasus-238M-Chinese
- Randeng-Pegasus-238M-Summary-Chinese
- Randeng-Pegasus-523M-Chinese
- Randeng-Pegasus-523M-Summary-Chinese

7.4.4 mT5
我们实现了中文版的mT5 (Xue et al., 2021)。为了加速训练，我们只在悟道语料库(180G版本)上重新训练了中英文对应的词汇和嵌入，并且使用了语料库自适应预训练(Corpus-Adaptive Pre-Training, CAPT)技术。
- Randeng-T5-77M
- Randeng-T5-784M

7.4.5 Transformer-Denoise
我们做了一个基于中文的Transformer模型的塔拉索实验，希望能够用其解决去噪任务。我们先使用Transformer-XL的模型结构在悟道语料库(180G版本)上进行预训练。然后在我们自主构建的去噪数据集上进行微调。其中，去噪任务是从包括随机插入/删除/替换/替换/句子重排的具有噪声的输入中重建一个流畅和干净的文本。
- Randeng-Transformer-1.1B-Denoise

7.5 太乙(MM)
为了处理多种模态，我们引入了太乙系列，并将其应用于跨模态场景。具体来说，我们正在研究蛋白质结构预测、语音-文本表示等任务。

7.5.1 CLIP
我们遵循CLIP (Radford et al., 2021)的实验设置，以获得强大的视觉-语言表征。在训练中文版的CLIP时，我们使用了chinese-robertawwm (Cui et al., 2021)作为语言的编码器，并将CLIP中的ViT应用于视觉的编码器。为了快速且稳定地进行预训练，我们冻结了视觉编码器并且只微调语言编码器。此外，我们将Noah-Wukong数据集 (Gu et al., 2022a)和Zero语料库 (Xie et al., 2022)用作预训练的数据集。据我们所知，我们的TaiyiCLIP是目前Huggingface社区中唯一的开源中文CLIP。
- Taiyi-CLIP-RoBERTa-326M-ViT-H-Chinese
- Taiyi-CLIP-RoBERTa-102M-Chinese
- Taiyi-CLIP-RoBERTa-large-326M-Chinese

7.5.2 MAP (暂定)
我们提出了一个探索性的多模态模型，以获得强大的表征。我们设计了一个特殊模块来解决多模态的不确定性问题。然后使用特殊的特殊训练策略对模型进行预训练。微调后的模型被应用于具有挑战性的多模态下游任务，并且能够实现实先的性能。我们的框架也有助于单模态的下游任务。其代码和相关预训练任务的详细信息将在论文被接收后公布。

并且，我们把Taiyi-CLIP模型应用于文本生成图像的场景中。其展示了强大的中文语言理解能力，比如可以利用中文古诗词生成图片。
- Taiyi-Roberta-124M-D
- Taiyi-Roberta-124M-D-v2
- Taiyi-vit-87M-D

7.6 余元(Domain)
尽管我们开源了很多解决通用需求的模型，但是在一些特定的领域中仍然有一些困难的问题尚未被解决。但领域特定的模型设计不同于一般的预训练语言模型，例如需要处理专有名词等。作为例子，在金融领域中，现有的模型如FinBERT (Yang et al., 2020)，它的词表和通用的BERT (Devlin et al., 2019)仅仅共

10https://github.com/fxsjy/jieba
7.6.1 BioBART
对于生物医学领域，我们采用了BioBART论文中介绍的BART (Yuan et al., 2022)结构。通过领域自适应预训练(Domain-Adaptive Pre-Training，DAPT)，我们把BART-large应用在PubMed摘要1上，其中包含大约41GB的生物医学研究论文的摘要。接着，我们收集各种生物医学语言的生成任务，包括对话、文本摘要、实体链接和命名实体识别，用作下游任务表现的评估。BioBART在不同的基准上都取得了很大的改进，并且与现有的最先进的方法相比，取得了具有竞争力或优越的结果。

- Yuyuan-Bart-139M
- Yuyuan-Bart-400M

7.6.2 GPT2
我们采用与Wenzhong-GPT2-3.5B相同的架构，在50GB的医学(PubMed)语料库上进行预训练。我们的Yuyuan-GPT2-3.5B是医疗领域最大的开源的GPT2模型。进一步地，模型可以通过计算困惑度(PPL)来判断事实。为了完成问答功能，我们将短语模式从疑问的形式转换为了陈述。

- Yuyuan-GPT2-3.5B
- YuyuanQA-GPT2-3.5B

7.7 系列名待定(探索)
我们也与其它组织一同进行一些实验性的探索。例如，以下模型是我们与追一科技共同努力的成功。

- Zhouwenwang-Unified-1.3B
- Zhouwenwang-Unified-110M

8 封神框架
为了解决第6节中提到的问题，我们结合了Huggingface (Wolf et al., 2019)，Megatron-LM (Shoeybi et al., 2019)，PyTorch-Lightning，和DeepSpeed (Rajbhandari et al., 2020)的优势并且融合进了封神框架中。如果用户熟悉上述框架，那么他们可以毫不费力地使用我们的深度学习框架。我们的框架支持对TB级数据，以及超过10亿的参数的大规模的模型进行预训练，并且支持对一系列的下游任务进行微调。通过一些配置，用户可以简单地进行分布式训练和使用节省内存的技术，让用户更加专注于模型部署和创新。封神框架还提供了我们在Huggingface中的开源的模型及应用的代码和示例。

我们的框架具有以下优势：
- 具有优于原始的Torch库的卓越性能，比如训练性能提升约300%。
- 支持大模型的模型：支持百亿级别内模型训练及微调。
- 支持超大规模的数据集(TB级)。
- 通过提供大量的预训练和下游任务的示例等，使得训练过程易于使用。
- 适应各种环境，比如支持在CPU，GPU，TPU等不同设备上运行。
- 集成主流的分布式训练逻辑，无需修改代码即可支持DDP和Zero Optimizer等分布式优化技术。

8.1 构架
封神框架中的核心组件如图5所示。我们的模型的核心功能对应图中的各个模块。如果在封神框架上进行开发，整体上可以按照下面的三个步骤进行：
1. 封装数据处理流程
2. 封装模型结构
3. 配置一些插件

8.2 文档
我们的文档12提供了分布式模型训练、微调以
及各种大模型应用的使用细节。此外，我们也展示了参与竞赛的论文以及各种可复现的代码和教程。

8.3 示例
如图6所示，在封神框架中构建模型只需要几行代码。对于已经在使用Huggingface之类的框架的用户来说，几乎不需要任何成本即可开始使用。

9 封神榜单
对于评估模型能力来说，好的基准是必要的。应该注意的是，中文在语言上与英文和其他印欧语系不同。因此，我们提出了封神榜单，它是由不同类型的中文自然语言任务组成的中文榜单。除了在任务中考虑公平的环境之外，我们还计划发布一些通用的预训练数据集。

9.1 任务选择的标准
为了构建高质量和健壮的基准，我们需要考虑到如何测试模型的方方面面。因此，我们在构建封神榜单时确定了以下要求：

广泛认可。虽然一些现有的数据集不是中文设计的，但它们多年来在NLP领域中被广泛使用，例如SuperGLUE (Wang et al., 2019)。所以，我们将召集一些专业的中英文的语言专家，精心翻译并校对这些热门的数据集。

面向未来。事实上，一些NLP模型已经在中国多个基准测试中超越了人类的表现。这宣告了人工智能已经拥有甚至可以超越人类水平的认知智能。我们认为的原因之一是这些基准的评估范围有限。更紧迫和必要是构建具有挑战性的数据集，而不是将现有数据集拉高到100%的准确度。未来的基准需要考虑更广泛的法官，技术和社会上的挑战。我们的数据集将会尽快发布，以更好地支持中文社区的进一步研究。

9.2 排行榜

9.2.1 Chinese-SuperGLUE
作为一个非常强大的基准，SuperGLUE被整个NLP社区广泛认可，但它仅提供英文版本。这些高质量的数据集也是中文社区所需要的。因为如果缺乏对英文数据集的验证，面向中文的模型可能会受到质疑。这是非常不公平的。为了分析中文数据集的质量，导致一些审稿人对一些没见过的数据集感到惊讶。为了解决上述问题，我们计划开发Chinese-SuperGLUE来评估中国模型。相关论文和榜单即将推出。

9.2.2 QAKM
为了评估NLP模型的知识水平，我们提出了QAKM（知识模型问答任务）。知识模型需要学习特定领域的知识并回答给定数据集中没见过的问题。该任务已包含在NLPCC13中，可在此网站14中访问。

10 总结和未来工作
本报告介绍我们的开源项目，封神榜，旨在成为中文认知智能的基础设施。封神榜的三个子项目（封神榜模型、封神框架、封神榜单）支持中文智能系统的各方面进展。此外，我们要强调的是，我们的封神榜项目是一个持续的开源项目，即我们会不断更新所有子项目。我们希望个人和组织的贡献者们也可以一起加入，共同完善该项目，共同构建整个中文社区。这个世界需要一些好的想法。

https://fengshenbang-doc.readthedocs.io/zh/latest/index.html

http://tcci.ccf.org.cn/conference/2022/cfpt.php

https://idea.edu.cn/ccnl-act/nlpcc-track1.html
伦理考量

我们的封神榜项目提供了整个中文生态系统来使用，生产和评估大模型，并且我们注意到许多公司和研究机构已经部署了我们的模型。我们的模型和基准正逐渐迁移到现实世界，从而对人类产生无法预测的影响。考虑伦理影响有很多方面需要关注：大规模模型的隐含偏见，潜在的环境问题，标记者语中的不良预判，开源框架的不当使用等。为了更好地理解并更深入了解讨论伦理问题，我们将在下一份报告中提供详细的讨论。我们鼓励任何开发者使用封神榜项目中的任何东西时就其使用展开公开辩论，例如任务如何选择和如何部署。我们希望这将减少任何不当行为发生的机会。

致谢

我们感谢CCNL的GTS团队提供的反馈和帮助。

References

Emily Alsentzer, John R. Murphy, Willie Boag, Weihung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. CoRR, abs/1904.03323.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In EMNLP/IJCNLP (1), pages 3613–3618. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. CoRR, abs/2004.05150.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kaixin Ethayarajh, Li Feifei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Kaeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. CoRR, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In NeurIPS.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. CoRR, abs/2010.02559.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotos Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen
Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese text encoder enhanced by n-gram representations. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 4729–4740. Association for Computational Linguistics.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. Codertext: A pre-trained model for programming and natural languages. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 1536–1547. Association for Computational Linguistics.

Yu Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022a. Wukong: 100 million large-scale Chinese cross-modal pre-training dataset and a foundation framework.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022b. Domain-specific language model pre-training for biomedical natural language processing. *ACM Trans. Comput. Hum.*, 3(1):2:1–2:23.

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufo, Shao Kun Deng, Colin B. Clement, Dawn Drain, Neel Sundaresan, Yuan Lin, Daxin Jiang, and Ming Zhou. 2021. Graphcodebert: Pre-training code representations with data flow. In *ICLR*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*. OpenReview.net.

L. Hsun. 2000. *A Brief History of Chinese Fiction*. University Press of the Pacific.
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In ICML, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67.

Samyang Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In SC, page 20. IEEE/ACM.

Mohammad Shoeybi, Mostofa Patwary, Raul Purz, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. CoRR, abs/1909.08053.

Yan Song, Tong Zhang, Yonggang Wang, and Kai-Fu Lee. 2021. ZEN 2.0: Continue training and adaptation for n-gram enhanced text encoders. CoRR, abs/2105.01279.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Reformer: Enhanced transformer with rotary position embedding. CoRR, abs/2104.09864.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In NeurIPS, pages 3261–3275.

Junjie Wang, Yuxiang Zhang, Ping Yang, and Ruyi Gan. 2022. Towards no.1 in CLUE semantic matching challenge: Pre-trained language model erlangsren with propensity-corrected loss. CoRR, abs/2208.02959.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. CoRR, abs/1910.03771.

Chunyu Xie, Heng Cai, Jianfei Song, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, Dawei Leng, Xiangyang Ji, and Yafeng Deng. 2022. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework.

Ruobin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanlan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In ICML, volume 119 of Proceedings of Machine Learning Research, pages 10524–10533. PMLR.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanli Li, Yina Patterson, Zuooyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2022. CLUE: A chinese language understanding evaluation benchmark. In COLING, pages 4762–4772. International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. M15: A massively multilingual pre-trained text-to-text transformer. In NAACL-HLT, pages 483–498. Association for Computational Linguistics.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. CoRR, abs/2006.08097.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yukuo Cen, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. In BioNLP@ACL, pages 97–109. Association for Computational Linguistics.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhihui Yang, and Jie Tang. 2021. WudaoCorpora: A super large-scale chinese corpora for pre-training language models. AI Open, 2:65–68.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In ICML, volume 119 of Proceedings of Machine Learning Research, pages 11328–11339. PMLR.

A List of Fengshenbang Models

This list is in alphabetical order.

1. Erlangshen-DeBERTa-v2-186M-Chinese-SentencePiece: https://huggingface.co/IDEA-CCNL/Erlangshen-DeBERTa-v2-186M-Chinese-SentencePiece
2. Erlangshen-DeBERTa-v2-320M-Chinese: https://huggingface.co/IDEA-CCNL/
   Erlangshen-DeBERTa-v2-320M-Chinese
3. Erlangshen-DeBERTa-v2-710M-Chinese: https://huggingface.co/IDEA-CCNL/
   Erlangshen-DeBERTa-v2-710M-Chinese
4. Erlangshen-DeBERTa-v2-97M-CWS-Chinese: https://huggingface.co/IDEA-CCNL/
   Erlangshen-DeBERTa-v2-97M-CWS-Chinese
5. Erlangshen-DeBERTa-v2-97M-Chinese: https://huggingface.co/IDEA-CCNL/
   Erlangshen-DeBERTa-v2-97M-Chinese
6. Erlangshen-Longformer-110M: https://huggingface.co/IDEA-CCNL/
   Erlangshen-Longformer-110M
7. Erlangshen-Longformer-330M: https://huggingface.co/IDEA-CCNL/
   Erlangshen-Longformer-330M
8. Erlangshen-MegatronBert-1.3B: https://huggingface.co/IDEA-CCNL/
   Erlangshen-MegatronBert-1.3B
9. Erlangshen-MegatronBert-1.3B-NLI: https://huggingface.co/IDEA-CCNL/
   Erlangshen-MegatronBert-1.3B-NLI
10. Erlangshen-MegatronBert-1.3B-Sentiment: https://huggingface.co/IDEA-CCNL/
    Erlangshen-MegatronBert-1.3B-Sentiment
11. Erlangshen-MegatronBert-1.3B-Similarity: https://huggingface.co/IDEA-CCNL/
    Erlangshen-MegatronBert-1.3B-Similarity
12. Erlangshen-MegatronBert-3.9B-Chinese: https://huggingface.co/IDEA-CCNL/
    Erlangshen-MegatronBert-3.9B-Chinese
13. Erlangshen-Roberta-110M-NLI: https://huggingface.co/IDEA-CCNL/
    Erlangshen-Roberta-110M-NLI
14. Erlangshen-Roberta-110M-Sentiment: https://huggingface.co/IDEA-CCNL/
    Erlangshen-Roberta-110M-Sentiment
15. Erlangshen-Roberta-110M-Similarity: https://huggingface.co/IDEA-CCNL/
    Erlangshen-Roberta-110M-Similarity
16. Erlangshen-Roberta-330M-NLI: https://huggingface.co/IDEA-CCNL/
    Erlangshen-Roberta-330M-NLI
17. Erlangshen-Roberta-330M-Sentiment: https://huggingface.co/IDEA-CCNL/
    Erlangshen-Roberta-330M-Sentiment
18. Erlangshen-Roberta-330M-Similarity: https://huggingface.co/IDEA-CCNL/
19. Erlangshen-Roberta-330M-Similarity
20. Erlangshen-UBert-110M-Chinese: https://huggingface.co/IDEA-CCNL/
    Erlangshen-UBert-110M-Chinese
21. Erlangshen-UBert-330M-Chinese: https://huggingface.co/IDEA-CCNL/
    Erlangshen-UBert-330M-Chinese
22. Erlangshen-ZEN1-224M-Chinese: https://huggingface.co/IDEA-CCNL/
    Erlangshen-ZEN1-224M-Chinese
23. Erlangshen-ZEN2-345M-Chinese: https://huggingface.co/IDEA-CCNL/
    Erlangshen-ZEN2-345M-Chinese
24. Randeng-BART-139M: https://huggingface.co/IDEA-CCNL/
    Randeng-BART-139M
25. Randeng-BART-139M-SUMMARY: https://huggingface.co/IDEA-CCNL/
    Randeng-BART-139M-SUMMARY
26. Randeng-BART-759M-Chinese-BertTokenizer: https://huggingface.co/IDEA-CCNL/
    Randeng-BART-759M-Chinese-BertTokenizer
27. Randeng-MegatronT5-770M: https://huggingface.co/IDEA-CCNL/
    Randeng-MegatronT5-770M
28. Randeng-Pegasus-238M-Chinese: https://huggingface.co/IDEA-CCNL/
    Randeng-Pegasus-238M-Chinese
29. Randeng-Pegasus-238M-Summary-Chinese: https://huggingface.co/IDEA-CCNL/
    Randeng-Pegasus-238M-Summary-Chinese
30. Randeng-Pegasus-523M-Chinese: https://huggingface.co/IDEA-CCNL/
    Randeng-Pegasus-523M-Chinese
31. Randeng-Pegasus-523M-Summary-Chinese: https://huggingface.co/IDEA-CCNL/
    Randeng-Pegasus-523M-Summary-Chinese
32. Randeng-T5-77M: https://huggingface.co/IDEA-CCNL/
    Randeng-T5-77M
33. Randeng-T5-784M: https://huggingface.co/IDEA-CCNL/
    Randeng-T5-784M
34. Randeng-Transformer-1.1B-Denoise: https://huggingface.co/IDEA-CCNL/
    Randeng-Transformer-1.1B-Denoise
35. Taiyi-CLIP-RoBERTa-326M-ViT-H-
B Author Contributions

Fengshenbang project is an ongoing open-source effort maintained by the team of engineers, researchers, interns at the Cognitive Computing and Natural Language (CCNL) Research Center in International Digital Economy Academy (IDEA).\footnote{https://www.idea.edu.cn/}

The project managers initiated the Fengshenbang project with its sub-projects, which includes:

- Jiaxing Zhang  
  \(\text{(zhangjiaxing@idea.edu.cn)}\)
- Ruyi Gan  
  \(\text{(ganruyi@idea.edu.cn)}\)

Other authors have contributed to the project in various ways, including data collection, designing and running experiments, constructing models, etc. Therefore, we provide a brief description:

- Fengshenbang Model: Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junyu Lu, Weifeng Chen, Junjie Wang, Junqing He, Yongfeng Huang, Xiayi Li, Yanghan Wu, Qi Yang, Yuxiang Zhang, Jianheng Zhuo, Xinyu Zhu, Zhongshen Zeng, Ting Han, Rui Wang, Xiaojun Wu, Kunhao Pan, Hao Wang, Chongpei Chen
- Fengshen Framework: Xinyu Gao, Ping Yang, Ziwei Wu
- Fengshenbang Benchmark: Junjie Wang, Ziwei Wu, Yuxiang Zhang
- Paper writing: Junjie Wang  
  \(\text{(wjj1020181822@toki.waseda.jp)}\), Yuxiang Zhang, Lin Zhang  \footnote{From CTO Lab in IDEA}