


PowerPulse: Power energy chat model with LLaMA model fine-tuned on Chinese and power sector domain knowledge

ChunLin Yin^{1,2} | KunPeng Du^{3,4} | Qiong Nong³ | HongCheng Zhang⁵ |
Li Yang¹ | Bin Yan⁵ | Xiang Huang¹ | XiaoBo Wang³ | Xuan Zhang^{3,6} 

¹Electric Power Research Institute of Yunnan Power Grid Co., Ltd., Kunming, China

²School of Information Science and Engineering, Yunnan University, Kunming, China

³School of Software, Yunnan University, Kunming, China

⁴School of Electromechanical Information, Yiwu Industrial & Commercial College, Jinhua, China

⁵Policy Research and Enterprise Management Department, Yunnan Power Grid Co., Ltd., Kunming, China

⁶Yunnan Key Laboratory of Software Engineering, Yunnan University, Kunming, China

Correspondence

Xuan Zhang, School of Software, Yunnan University, Kunming, Yunnan 650091 China.
Email: zhxuan@ynu.edu.cn

Funding information

the Major Project of Yunnan Natural Science Foundation, Grant/Award Number: 202302AE09002003; the Open Foundation of Yunnan Key Laboratory of Software Engineering, Grant/Award Number: 2023SE101; the Science and Technology Project of Yunnan Power Grid Co., Ltd, Grant/Award Number: YNKJXM20222254; the Science Foundation of "Knowledge-Driven Intelligent Software Engineering Innovation Team"; the Science Foundation of Young and Middle-aged Academic and Technical Leaders of Yunnan, Grant/Award Number: 202205AC160040; the Science Foundation of Yunnan Jinzhi Expert Workstation, Grant/Award Number: 202205AF150006

Abstract

Recently, large-scale language models (LLMs) such as chat generative pre-trained transformer and generative pre-trained transformer 4 have demonstrated remarkable performance in the general domain. However, inadaptability in a particular domain has led to hallucination for these LLMs when responding in specific domain contexts. The issue has attracted widespread attention, existing domain-centered fine-tuning efforts have predominantly focused on sectors like medical, financial, and legal, leaving critical areas such as power energy relatively unexplored. To bridge this gap, this paper introduces a novel power energy chat model called PowerPulse. Built upon the open and efficient foundation language models (LLaMA) architecture, PowerPulse is fine-tuned specifically on Chinese Power Sector Domain Knowledge. This work marks the inaugural application of the LLaMA model in the field of power energy. By leveraging pertinent pre-training data and instruction fine-tuning datasets tailored for the power energy domain, the PowerPulse model showcases exceptional performance in tasks such as text generation, summary extraction, and topic classification. Experimental results validate the efficacy of the PowerPulse model, making significant contributions to the advancement of specialized language models in specific domains.

KEYWORDS

domain-specific knowledge, instruction fine-tuning, LLaMA, LLMs, LoRA, power energy

1 | INTRODUCTION

In recent years, large-scale language models (LLMs) have garnered significant attention in natural language processing (NLP), with models such as OpenAI's chat generative pre-trained transformer (ChatGPT) (Ouyang et al., 2022) and generative pre-trained transformer 4 (GPT-4)

ChunLin Yin, KunPeng Du, and Qiong Nong contributed equally to this work and should be considered co-first authors.

(OpenAI, 2023) making remarkable strides in understanding and generating human-like text. These models leverage extensive pre-training on unlabeled natural language data to grasp grammar, semantics, and context, enhancing their ability to comprehend complex instructions (Yunxiang et al., 2023).

While instruction-based LLMs excel in generating human-like responses in dynamic dialogue environments, they face challenges when applied to specialized domains like power energy, where specific domain knowledge is critical for accurate text processing (Cui et al., 2023). Generic LLMs may generate responses that lack domain-specific accuracy, leading to hallucination issues (Ji et al., 2023a; Li et al., 2022).

The power energy domain encompasses a wide array of knowledge, including energy structures, institutional information, power theories, and more. Accurate comprehension and integration of this domain-specific knowledge are crucial for tasks such as decision-making and predictive analytics. This necessitates the development of domain-specific language models. In response to this need, we introduce PowerPulse, a specialized LLM for the power energy domain. Built upon the open-source LLaMA framework (Touvron et al., 2023), PowerPulse undergoes incremental pre-training on 4 million high-quality Chinese power energy texts. Additionally, a knowledge graph for the power energy domain is designed to provide structured information. A fine-tuning dataset of 200 k instruction data points further enhances the model's domain-specific understanding.

PowerPulse demonstrates outstanding performance in tasks such as text generation, summary extraction, and topic classification, providing a valuable solution for domain-specific language modelling. Our contributions include:

1. This paper presents the first attempt to train a Chinese LLM specifically for the field of power energy, improving Chinese power knowledge understanding capabilities of LLM through an additional collection of 4 million high-quality Chinese power text data.
2. A new dataset is created containing 200k instruction data points for fine-tuning LLM, which includes a substantial amount of domain-specific knowledge in the field of power energy.
3. We also provide benchmarks for evaluating the performance of LLMs in instruction tracking tasks and natural language understanding tasks in the field of power energy.
4. The research resources and outcomes of this paper can promote further research and collaboration in the power energy domain, encouraging power energy institutions to easily train their own LLMs based on internal data.

In conclusion, PowerPulse represents a significant step in addressing the unique challenges of domain-specific language modelling in the power energy sector.

The structure of our study is as follows: Section 2 provides an overview of related research. Section 3 details the framework of our PowerPulse model, encompassing dataset construction and instruction development. In Section 4, we elucidate our experimental procedures, including the setup, and conduct an in-depth analysis of the obtained results. Our findings are synthesized and avenues for future research are explored in Section 5.

2 | RELATED WORK

2.1 | Domain-specific LLMs

Ever since the advent of the transformer architecture (Vaswani et al., 2017), LLMs have made significant strides in the field of NLP. Prior to the emergence of ChatGPT, researchers primarily relied on bidirectional encoder pre-training models such as BERT (Devlin et al., 2018) and Roberta (Liu et al., 2019), which excelled in general NLP tasks across different domains. Furthermore, domain-specific variants of BERT, such as BioBERT (Lee et al., 2020), NukeLM (Burke et al., 2021), NukeBERT (Jain et al., 2020), and MatSciBERT (Gupta et al., 2022), achieved important breakthroughs in tasks related to natural language understanding in domains such as biology, energy, and finance. These domain-specific LLMs, pre-trained and fine-tuned on specific domain data, effectively captured semantic and contextual information, providing powerful NLP tools for domain experts.

Recent advancements in LLMs indicate that instruction-based autoregressive language models outperform the previous generation of bidirectional encoder pre-training models. Notably, the significant increase in model scale has brought about emerging abilities, including contextual learning for zero-shot tasks and enhanced performance on complex tasks (Touvron et al., 2023; Wu et al., 2023). Particularly, the introduction of OpenAI's ChatGPT and GPT-4 caused a sensation across various industries, leading to a paradigm shift in how LLMs are perceived. However, OpenAI has not publicly disclosed the detailed training strategies or weight parameters of ChatGPT, and such generic LLMs may lack explicit design for specific domains, potentially lacking adaptability to domain-specific terminology, language style, contextual information, and task requirements. As a result, researchers have turned their attention to training and applying domain-specific LLMs, such as BloombergGPT (Wu et al., 2023), FinGPT (Yang et al., 2023), PIXIU (Xie et al., 2023) for the financial sector; ChatDoctor (Yunxiang et al., 2023), DoctorGLM (Xiong et al., 2023), HuaTuo (Wang et al., 2023), PaLM (Xiong et al., 2023), LLaVA-Med (Li et al., 2023a), Med-PaLM 2 (Singhal et al., 2023) for the

medical field; Lawyer-LLaMA (Huang et al., 2023), LawGPT (Gentile, 2023) for the legal field; and LLMs (Taylor et al., 2022) for the scientific domain, as well as LLMs (Chen et al., 2021) for code-related domain. These domain-specific LLMs have achieved significant success through fine-tuning LLMs available in the open-source community.

Research in the field of electric power and energy, being an essential foundational industry in modern society, has been relatively limited. This is often due to the vast and complex textual data present in this domain, ranging from energy policies and regulations to power equipment maintenance records, encompassing a wide variety of information. Additionally, tasks in the field of power energy are diverse, involving tasks such as text generation, comprehension, classification, and dialogue, with a lack of available corpora and relevant instruction datasets. Despite facing a series of challenges, the prospects of training LLMs in the field of power energy remain promising. By further refining pre-training and fine-tuning techniques that leverage instruction guidance for LLMs and incorporating domain expertise, the performance of models in this domain can be enhanced. This will bring numerous benefits to the energy and power industries, including improved efficiency in the energy sector, intelligent management of power equipment, superior user service experiences, and paving the way for power energy institutions to train their own LLMs based on internal data.

2.2 | Low-rank adaptation

In the field of NLP, fine-tuning LLMs for specific domains and tasks has always posed a significant challenge. As model sizes continue to grow, fine-tuning all parameters of the model has become impractical in terms of both cost and time for research laboratories and companies. To tackle this issue, researchers have proposed several methods, such as adapters (Houlsby et al., 2019) and prefix-tuning (Li & Liang, 2021). However, adapters introduce additional inference latency as the model depth increases, while prefix-tuning is relatively difficult to train and may not be as effective as direct fine-tuning.

Studies by Aghajanyan et al. (2021) have demonstrated that the learned over-parameterized models in fact reside on a low intrinsic dimension. Drawing inspiration from these studies, Hu et al. (2021) put forth Low-Rank Adaptation, called LoRA. LoRA assumes that the weight changes during the adaptation process have a low 'intrinsic rank', enabling the model to indirectly train some dense layers in the neural network by optimizing a low-rank decomposition matrix, while keeping the pre-trained weights unchanged. This approach allows freezing the pre-trained model weights and injecting trainable low-rank decomposition matrices into each layer of the model, significantly reducing the number of trainable parameters for downstream tasks. Consequently, this reduces computational complexity and memory requirements. LoRA provides an efficient solution for lightweight fine-tuning and effectively approximates full model parameter fine-tuning, enabling fine-tuning with limited computational resources while maintaining model performance. This approach substantially reduces the total number of trainable parameters, making it feasible to train LLMs with fewer computational resources (Cui et al., 2023).

2.3 | Instruction learning

Instruction learning, also known as instruction fine-tuning, is a method that enables LLMs to perform tasks based on specific instructions. It aims to make LLMs respond accurately and reliably to specific questions or descriptive instructions, rather than just having a general understanding of natural language. Instruction learning is often achieved through instruction fine-tuning, where LLMs receive descriptive instructions or human requests during the training process to generalize to natural interactions in new scenarios. The training process of instruction learning typically employs supervised learning or reinforcement learning methods. In the supervised learning (Sanh et al., 2022; Wei et al., 2021), LLMs receive example instructions from the instruction dataset and are fine-tuned based on the provided sample outputs. In the reinforcement learning (Minsky, 1961), LLMs adjust their model parameters based on environmental feedback or reward signals to achieve better task execution results (Ouyang et al., 2022).

The key elements of instruction learning include diverse and representative instruction datasets to ensure that LLMs can effectively perform tasks in a wide range of scenarios and possess generalization capabilities. The instruction dataset should cover various real-world problems and encompass diverse language styles and expressions. Researchers have already created instruction datasets in the general domain, including general dialogue-based instruction datasets (Ji et al., 2023b; Peng et al., 2023; Zhang et al., 2023). These datasets serve as the foundation for LLMs' instruction fine-tuning, enabling the models to accept instructions and perform corresponding tasks, leading to significant performance improvements in the general domain. However, specific domain instruction datasets (Ding et al., 2023; Li et al., 2023b), especially in industries like power and energy, are relatively scarce. There is a lack of available domain fine-tuning instruction datasets, posing challenges to the effective application of instruction learning. Therefore, further research is needed to construct representative instruction datasets to enhance LLMs' generalization capabilities, enabling them to intelligently execute power and energy-related tasks, thus improving energy industry efficiency and enhancing user service experiences.

2.4 | Hallucination issues

In the realm of NLP and text generation models, 'hallucination' refers to the generation of content in the model's output that lacks logical coherence or factual accuracy, and cannot be substantiated by input information. There are diverse causes for the occurrence of hallucination issues in text generation models (Ji et al., 2023a).

The most conspicuous factor is the disparities between input and output data. When discrepancies arise between the input-output pairs in the training dataset, models tend to generate hallucinated content during the training process. From a data perspective, a fundamental cause is unguided generation. In tasks like 'data-to-text', when the generated text extends beyond the scope of the input data, the model indulges in unbridled creative writing, leading to hallucination. At the task level, inherent disparities between input and output, such as in dialogue tasks relying on prompt-based inputs, can also contribute to hallucination. Furthermore, exposure bias, stemming from the discrepancy between training and inference, can lead to hallucination. Training relies on the true input, whereas inference depends on previously generated model output, introducing inconsistencies and potential hallucination. Additionally, large-scale pre-trained language models introduce inherent 'knowledge' parameters that do not exist in the input data, further increasing the likelihood of hallucination (Ye et al., 2023).

In the context of the power energy domain, where precise understanding and application of domain knowledge are paramount, PowerPulse provides a solution to address hallucination issues. It is specifically designed to mitigate hallucination by undergoing incremental pre-training on a substantial corpus of high-quality Chinese power energy texts. Additionally, a domain-specific knowledge graph is constructed to provide structured, reliable information, reducing the likelihood of hallucinated content. Fine-tuning on a dataset containing 200 k instruction data points further enhances PowerPulse's domain-specific understanding, enabling it to generate accurate and contextually relevant responses in the power energy domain.

3 | PROPOSED METHOD

In this section, we commence by presenting the Base Model in Section 3.1, followed by a comprehensive exposition of the data collection methodology and the construction of the instruction dataset in Section 3.2. Finally, in Section 3.3, an in-depth account of the training process for PowerPulse is provided. The overall architectural of our proposed PowerPulse is illustrated in Figure 1.

3.1 | Base model

LLaMA (Touvron et al., 2023) is an open and efficient foundation LLM exclusively designed for decoder tasks, built upon the transformer architecture (Vaswani et al., 2017), and constitutes an amalgamation of multilingual base models. It comprises embedding layers, multiple transformer blocks, and a language modelling head. Currently, the model offers four different size options: 7B, 13B, 33B, and 65B, and has been open-sourced to the research community. During pre-training, LLaMA leverages a diverse range of publicly available resources, including web crawls, books, Wikipedia, and preprint papers. These comprehensive pre-training data contribute to enhancing the model's performance and adaptability. In this paper, we adopted the Chinese Alpaca Plus 7B Model, made accessible by Xu et al. (2023), for the purpose of more convenient training. This model is built upon the Meta original LLaMA (Touvron et al., 2023) and involves multiple LoRA weight merging processes. Specifically, to enrich the training data, Xu et al. (2023) first introduced a Chinese vocabulary and continued the pre-training process, expanding LLaMA to 120G

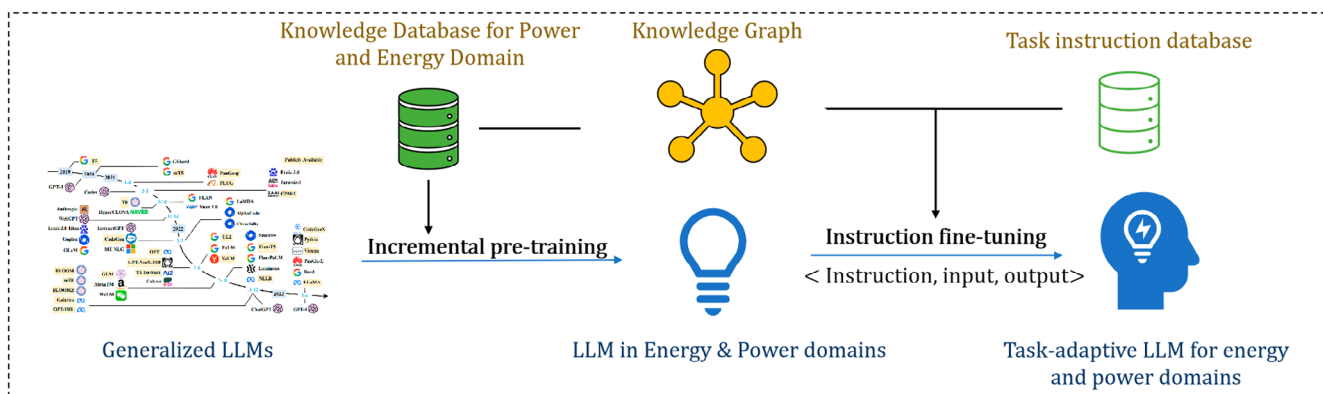


FIGURE 1 Framework of PowerPulse.

of generic domain text and Alpaca to 4 M instruction data, with a particular focus on incorporating STEM-related data. Subsequently, the model weights of the Chinese LLaMA LoRA and Chinese Alpaca LoRA were merged, resulting in the creation of the Chinese Alpaca Plus 7B Model. This model is a publicly accessible Chinese general LLMs that supports continued training.

The choice of the Chinese Alpaca Plus 7B Model as the foundation for PowerPulse is twofold. Firstly, LLaMA contains only 7 billion parameters, yet it has demonstrated outstanding performance and competitive capabilities compared to much larger models like GPT-3 (Brown et al., 2020), which has 175 billion parameters, across multiple NLP benchmarks. Its significantly reduced parameter size makes it a more feasible option for research labs or companies in terms of cost and time. Secondly, LLaMA's training set consists of approximately 1.4 T tokens, with a majority being in English (Cui et al., 2023). As a result, LLaMA has limitations in generating Chinese text. The Chinese Alpaca Plus 7B Model addresses this limitation by leveraging Chinese language corpora to expand its Chinese vocabulary, thereby enhancing LLaMA's comprehension and generation abilities for Chinese text. Overall, the utilization of the Chinese Alpaca Plus 7B Model as the foundation for PowerPulse not only allows for more accessible and efficient training but also enhances the model's performance in generating Chinese contents. This combination of advantages makes the model a suitable choice for the intended application in the field of power energy.

3.2 | Power and energy-based data

3.2.1 | Building a knowledge base for power energy domain

In this paper, we first collected data for incremental pre-training, with the aim of conducting secondary pre-training of LLMs on a massive corpus of domain-specific documents to inject domain knowledge. To gather data in the field of power energy, a variety of methods were employed. Initially, 1.6 million Internet public documents were crawled from over 500 websites and 1000 URL addresses, providing a wealth of information on policies, regulations, technological developments, and market trends in the power energy industry. Additionally, we monitored multiple WeChat public accounts belonging to experts and institutions in the power energy industry, obtaining real-time updates on policy dynamics, technological advancements, and market trends, allowing us to stay abreast of industry developments and policy changes. Finally, we collected relevant news, policy documents, and announcements from authoritative public websites such as the National Development and Reform Commission, the National Energy Administration, and the China Electric Power Network, ensuring the comprehensiveness and accuracy of the data.

Regarding data processing, the policy-related information was first cleaned by removing HTML tags, extracting key details, and standardizing date and time formats for ease of subsequent analysis. Then, the collected data was categorized and organized based on different categories, including policies and regulations, market dynamics, technological researches, and corporate activities. Some samples are shown in Table 1, laying the foundation for further analysis and utilization. This comprehensive and detailed data collection and processing process provided a reliable data basis for our work in this paper.

3.2.2 | Power and energy knowledge graph

As structured data, the power energy knowledge graph encompass data related to entity categories and attributes, including energy knowledge, institutional knowledge, power theory knowledge, power equipment knowledge, meteorological disaster knowledge, and electricity usage knowledge. Additionally, professional data on experts' research achievements, paper information, and collaboration status was also collected. However, due to its structured nature, the data is not directly suitable for incremental pre-training of LLMs. Inspired by self-instruct construction (Wang et al., 2022), we treated the entity relationships within the constructed knowledge graph as keywords and leveraged the ChatGPT API to generate a large volume of high-quality textual data specific to the power energy industry. Finally, some instances in the knowledge base for power energy industries are shown in Table 1.

3.2.3 | Dataset creation for instructional fine-tuning

As we need a set of instructions specific to the field of power energy to fine-tune LLMs for handling various tasks within the domain, following the approach of self-instruct (Wang et al., 2022), the instruction dataset for the field of power energy was generated. As shown in Figure 2, for the question answering in the power energy contexts, the database was curated and organized from the knowledge base on power energy industry. This database can be updated at any time without requiring model retraining, potentially tailored to different energy sectors or specific objectives. Emphasis was placed on tasks such as text generation, comprehension, classification, and dialogue, aiming to meet the basic requirements of the power energy industry while seeking performance improvements in relevant domain tasks and providing support for decision-making and policy research.

TABLE 1 Instances in the knowledge base for power energy industries.

Categories	Instances (in Chinese)	Instances (translated to English)
Enterprise dynamics	<p><浙江首个多源协同山区配电网在泰顺启用> 8月16日10时，在国网浙江泰顺供电公司调度大厅，伴随着调度员一条条复令回响，浙江首个多源协同山区配电网在温州泰顺投产启用。泰顺地处浙江南部山区，境内90%被群山覆盖，存在电网结构弱、抗干扰能力低等问题。位于泰顺北部山脉的南浦溪镇，20余个乡村仅依靠一条50千米长的线路来保证用电，在遭遇台风、雷击等自然灾害时，山区居民的用电可靠性难以得到保障。如果在山区建设一座35千伏变电站，则要花费近5000万元的成本，建设难度大、周期长、可行性低。—</p>	<p><Zhejiang's first multi-source synergistic mountain power distribution network in Taishun opened.> At 10 o'clock on August 16, in the State Grid Zhejiang Taishun Power Supply Company scheduling hall, accompanied by the dispatcher of an order echo, Zhejiang's first multi-source synergistic mountain distribution network in Taishun, Wenzhou, put into operation. Taishun is in the southern mountainous areas of Zhejiang, 90% of the territory is covered by mountains, there is a weak grid structure, low interference resistance and other issues. Located in the northern mountains of Taishun South Puxi town, more than 20 villages rely only on a 50-kilometer-long line to ensure that electricity, in the event of a typhoon, lightning strikes and other natural disasters, the reliability of the power of the residents of the mountainous areas is difficult to be guaranteed. —</p>
Policies and regulations	<p><全球首份数字电网实践白皮书发布> 11月13日，第十七届中国南方电网国际技术论坛暨《数字电网白皮书》发布会在深圳前海万科国际会议中心举行。会上，南方电网发布全球第一份《数字电网白皮书》—</p>	<p><World's First White Paper on Digital Grid Practices Released> On November 13th, the 17th China Southern Power Grid International Technology Forum and the release ceremony of the "Digital Grid White Paper" were held at the Shenzhen Qianhai Vanke International Conference Center. During the event, China Southern Power Grid unveiled the world's first "Digital Grid White Paper."—</p>
Power energy knowledge graph	<p>以下文本中的实体关系三元组如下:<电力网, 包含, 变电所、输电线路、配电线路><电力网, 功能, 输送与分配电能> 电力系统中各种电压的变电所及输配电线路组成的整体, 称为电力网。它包含变电、输电、配电三个单元。电力网的任务是输送与分配电能, 改变电压。</p>	<p>The entity relation triples in the given text are as follows:<Power Grid, Contains, Substations, transmission lines, distribution lines> and <Power Grid, Function, Transmission and distribution of electricity> The entirety composed of substations for various voltage levels and transmission and distribution lines in the power system is referred to as the power grid. It consists of three units: substations, transmission, and distribution. The power grid's task is to transport and distribute electricity, and to alter voltage.</p>

The framework for collecting and structuring power domain data is shown in Figure 2. A dataset of 4,329,891 texts was collected for incremental pre-training of LLMs, along with 200k instruction data used as training instances for supervised fine-tuning.

3.3 | Training of PowerPulse

The Chinese Alpaca Plus 7B Model, open-sourced by Xu et al. (2023), is adopted as the base model for PowerPulse. To further enhance PowerPulse's adaptability to the terminology, language style, and task requirements in the field of power energy, a domain-specific knowledge base consisting of 4,329,891 high-quality Chinese texts related to power energy is utilized for incremental pre-training in Section 3.2. This incremental pre-training enables PowerPulse to specialize in the domain and acquire profound knowledge.

Then, we fine-tune PowerPulse using LoRA with only 7 million trainable parameters. The fine-tuning process utilizes a dialogue dataset comprising 200 k instructions and is conducted on 4 * 3090Ti GPUs. The hyperparameters employed during training are as follows: a batch size of 4, a learning rate of 2e-5 for the Lion optimizer, a total of 5 epochs, a maximum sequence length of 512 tokens, and a maximum target length of 200 tokens.

$$h = W_0x + \Delta Wx = (W_0 + BA)x, B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}. \quad (1)$$

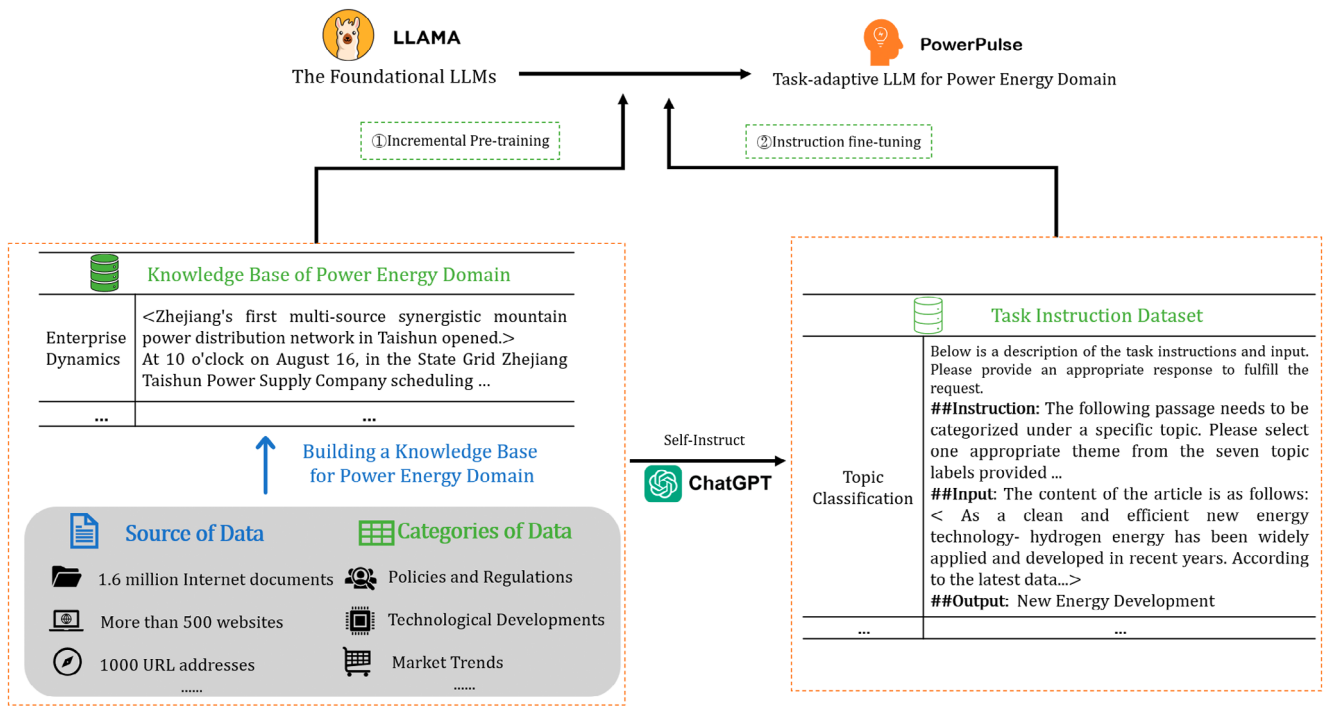


FIGURE 2 Collecting and structuring power domain data. The dashed box on the left represents the data source of the knowledge base for the power energy domain and the compositional structure of this knowledge base. The dashed box on the right provides an instructional example for topic classification tasks constructed through self-instruct method based on the data within the knowledge base. In addition, each instruction sample follows a ‘description-instruction-input-output’ structure.

Specifically, for the PowerPulse model, the initial parameter matrix is denoted as $W_0 \in \mathbb{R}^{d \times k}$, where k represents the input dimension, and d is the output dimension. The parameter matrix ΔW is learned during the fine-tuning process of PowerPulse. LoRA assumes that ΔW is a low-rank matrix. To further decompose ΔW , we express it as the product of two trainable matrices, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, where r is a predetermined hyperparameter, and $r < \min(d, k)$. During the training process, the initial parameters W_0 are fixed and frozen, meaning they do not receive gradient updates. Instead, only the parameters of the smaller matrices B and A are updated.

When deploying the model in a production environment, one can simply merge the initial parameter matrix W_0 with the LoRA parameters to obtain the fine-tuned model parameters: $W = (W_0 + BA)$. The incorporation of LoRA does not introduce inference overhead, and the inference process can be carried out in the same way as before, without introducing additional latency. Specifically, prior to fine-tuning, the computation of h is done as $h = W_0x$, and after fine-tuning, it becomes $h = Wx$, with no extra delay introduced.

In summary, the proposed LoRA method enables efficient fine-tuning of the PowerPulse model by decomposing the parameter matrix ΔW into smaller trainable matrices, thus significantly reducing the number of fine-tuning parameters. The seamless merging of initial and fine-tuned model parameters allows for smooth deployment without any inference overhead.

4 | EXPERIMENTS

In this section, we assess the performance of PowerPulse in tasks such as text generation, summary extraction, and topic classification. Firstly, in Section 4.1, we present the setup of the experimental datasets and evaluation metrics. Subsequently, in Section 4.2, we introduce the experimental baselines and compare PowerPulse against state-of-the-art LLMs to comprehensively evaluate its capabilities.

4.1 | Experimental settings

4.1.1 | Dataset

We constructed a dataset named PowerCorp for evaluating the performance of LLMs in the field of power energy. Our main tasks involve text generation, summary extraction, and topic classification. Carefully curated, the dataset comprises a challenging set of samples that demand the

model to produce high-quality, coherent text within the field of power energy. Table 2 shows the experimental settings for the PowerCorp dataset.

In the text generation task, our aim is to assess the model's ability to produce relevant text expressions related to energy and power industries. This portion of the dataset includes text data gathered from various sources of power energy literature, encompassing diverse language styles and text types.

The summary extraction task requires the model to automatically extract key information and points related to the energy and power knowledge from a large corpus of text. This part of the dataset contains a diverse set of power energy documents, including academic papers, news reports, government documents, and so forth.

Text classification tasks represent a crucial facet of the field of NLP, with extensive applicability across various domains. These applications encompass sentiment analysis (Ali et al., 2023; Rahman & Halim, 2023; Tahir et al., 2023), spam detection (Halim et al., 2020), paraphrase identification (Du et al., 2023), topic classification among others. The primary objective of text classification is to allocate textual documents to predefined categories or labels. Among other things, we focus on the topic categorization task, which evaluate whether the model can distinguish and comprehend subtle differences between different topics, showcasing its performance in natural language understanding. The dataset for this task includes seven topic labels, that is, energy price change, power supply and demand, grid planning, international energy cooperation, energy market trend, new power system, energy policy report.

4.1.2 | Evaluation Metrics

To automatically assess the readability of texts generated by LLMs based on given prompts and identify potential substantive differences from texts written by experts, we have opted for specific metrics. We employ readability metric GFI to measure the comprehensibility of texts, and ROUGE to evaluate the quality of the generated content.

The Gunning Fog Index (GFI) (DuBay, 2004) is one such readability metric used to test the readability of writing. It is commonly utilized to ascertain whether the target audience can easily comprehend the text. The formula for its calculation is as follows:

$$GFI = \alpha \times [ASL + \beta \times ACW]. \quad (2)$$

The parameters α and β are randomly initialized values in the experiment, which are set to 0.5 and 1, respectively. ASL represents the average sentence length in one or more complete paragraphs, calculated by dividing the total number of words by the number of sentences. ACW represents the proportion of 'complex' words in the total word count. In the Chinese context, this is determined by the ratio of adverbs and conjunctions in each sentence.

TABLE 2 Experimental settings for the PowerCorp dataset.

Tasks	Instructions (in Chinese)	Instructions (translated to English)
Text Generation	下面是一个描述任务的指令和输入。请用适当的回答完成请求。 ### Instruction:请以给定的题目，写一篇内容500字左右的精彩报道。 ### Input:题目为:<>	Below is a description of the task instructions and input. Please provide an appropriate response to fulfill the request. ### Instruction: Compose a captivating article of approximately 500 words based on the given topic. ### Input: Topic is <>
Summary Extraction	下面是一个描述任务的指令和输入。请用适当的回答完成请求。 ### Instruction:以下文章需要一个合适的标题，以清晰概括其核心内容。 ### Input:文章内容如下:<>	Below is a description of the task instructions and input. Please provide an appropriate response to fulfill the request. ### Instruction: The following text requires a suitable title to concisely summarize its core content. ### Input: The content of the text is as follows: <>
Topic Classification	下面是一个描述任务的指令和输入。请用适当的回答完成请求。 ### Instruction:以下文章需要进行主题分类，请从以下7个主题标签中选择一个作为合适的主题，主题标签包括:'能源价格变化'、'电力供需'、'电网规划'、'能源国际合作'、'能源市场动态'、'新型电力系统'、'能源政策报告'。 ### Input:文章内容如下:<>	Below is a description of the task instructions and input. Please provide an appropriate response to fulfill the request. ### Instruction: The following passage needs to be categorized under a specific topic. Please select one appropriate theme from the seven topic labels provided: 'Energy Price Change', 'Electricity Supply and Demand', 'Grid Planning', 'International Energy Cooperation', 'Energy Market Trend', 'New Power System', 'Energy Policy Report'. ### Input: The content of the article is as follows: <>

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is an evaluation metric used to assess the quality of text summaries. It is widely applied in text generation tasks, particularly in the field of text summarization. It evaluates the similarity between the system-generated summary and the reference summary. This is achieved by measuring the overlap of n -grams (continuous sequences of n words) between the generated summary and the reference summary. The calculation formula is as follows:

$$\text{ROUGE}_n = \frac{\sum_{r \in R} \sum_{g \in G} \text{lcs}_n(r, g)}{\sum_{r \in R} \sum_{g \in G} (|r| + 1 - n) [\text{lcs}_n(r, g) > 0]}, \quad (3)$$

where R refers to the set of reference answers, and G represents the set of generated results. $\text{lcs}_n(r, g)$ denotes the length of the longest common subsequence of n -grams between the reference answers and generated results.

The ROUGE-L metric calculates the degree of sentence completeness based on the longest common sequence of the two text units to identify the degree of coverage of duplicate words and length sequences in the abstracts versus the reference abstracts with the following equation:

$$L(x, y, l) = \frac{\text{LCS}(x, y)}{l}, \quad (4)$$

where $\text{LCS}(x, y)$ is the longest common sequence of two text units x , and y , l is the length of the text.

$$\text{ROUGE-L} = \frac{(1 + \beta^2) L(x, y, m) \times L(x, y, n)}{\beta^2 L(x, y, m) + L(x, y, n)}, \quad (5)$$

where x is the reference summary, with the length of m ; y is the generated summary, of length n ; β is the ratio of precision to recall.

Accuracy is a measure of precision, representing the proportion of correctly judged samples out of all samples.

4.2 | Baseline models

Chinese-LLaMA (Cui et al., 2023; Xu et al., 2023): as the foundational model for PowerPulse, is a LLaMA (Touvron et al., 2023) designed specifically for decoder tasks. It is built upon the transformer (Vaswani et al., 2017) architecture and constitutes a collection of multilingual base models. LLaMA consists of embedding layers, multiple transformer blocks, and a language modelling head. In this paper, Chinese-LLaMA-7B (Cui et al., 2023; Xu et al., 2023) was chosen for the Chinese language environment, considering its adoption from the perspective of fairness and training cost. It can be accessed through the following link <https://github.com/ymcui/Chinese-LLaMA-Alpaca/tree/v3.0>.

ChatGLM2 (Zeng et al., 2022): ChatGLM2 is the second-generation version of the open-source bilingual dialogue model ChatGLM-6B. It retains numerous outstanding features of ChatGLM-6B, such as smooth conversation flow and low deployment threshold, while also supporting longer contexts, more efficient inference, and improved performance.

Bloomz (Muennighoff et al., 2022): Bloomz is able to follow models of human instructions in dozens of languages in zero-shot scenarios. It fine-tunes pre-trained multilingual language models on cross-linguistic tasks and finds that the resulting models generalize cross-linguistically to unseen tasks and languages.

ChatGPT (Ouyang et al., 2022): ChatGPT, a state-of-the-art conversational AI model, is a variant of the GPT architecture fine-tuned for natural language understanding and generation in dialogue settings. It excels in generating human-like responses, making it valuable for chatbots, virtual assistants, and interactive applications.

4.3 | Experimental results

In this subsection, the experimental baselines are presented and PowerPulse is compared with them to evaluate its performances.

4.3.1 | Main result

Table 3 presents the GFI values of PowerPulse and other state-of-the-art LLMs in the text generation task, evaluated under both zero-shot and 1-shot settings. The GFI values are used to assess the readability of text generated by LLMs, where higher values indicate higher complexity and poorer readability. Our model PowerPulse achieves the best results, indicated in bold. The second-best results are underlined. Therefore, in the text generation task, our PowerPulse demonstrates superior performance.

TABLE 3 GFI values of LLMs in text generation task under zero and 1-shot settings.

Models	Zero-shot	1-shot
LLaMA (Touvron et al., 2023)	1.5334	1.5661
ChatGLM2 (Du et al., 2021)	1.5372	1.4411
Bloomz (Muennighoff et al., 2022)	2.9808	7.1954
ChatGPT (Ouyang et al., 2022)	<u>1.4456</u>	<u>1.3955</u>
PowerPulse (Ours)	1.1410	1.2380

TABLE 4 ROUGE values of LLMs in summary extraction task under zero and 1-shot settings.

Models	Zero-shot		1-shot			
	ROUGE -1	ROUGE -2	ROUGE -L	ROUGE -1	ROUGE -2	ROUGE -L
LLaMA (Touvron et al., 2023)	0.1618	0.0701	0.1468	0.0834	0.0374	0.0779
ChatGLM2 (Du et al., 2021)	0.1989	0.0864	0.1796	0.1781	0.0809	0.1573
Bloomz (Muennighoff et al., 2022)	0.0887	0.034	0.0825	0.0163	0.0120	0.0159
ChatGPT (Ouyang et al., 2022)	<u>0.3464</u>	0.2815	<u>0.3152</u>	0.3285	0.1675	<u>0.2273</u>
PowerPulse (Ours)	0.3945	<u>0.2504</u>	0.3539	<u>0.2902</u>	<u>0.1501</u>	0.2373

As shown in Table 3, PowerPulse achieves GFI values of 1.1410 and 1.2380 under the zero-shot and 1-shot settings, respectively, significantly lower than those of other LLMs. This indicates that PowerPulse performs well in terms of complexity and readability of the generated text, approaching human-level text generation. Consequently, PowerPulse exhibits higher complexity and stronger generation capabilities, making it more suitable for chat tasks in the field of power energy. In contrast, LLaMA (Touvron et al., 2023) achieves a GFI value of 1.5334 under the zero-shot setting, second only to PowerPulse. This demonstrates that LLaMA (Touvron et al., 2023) possesses a certain level of competence in the text generation task, even without domain-specific fine-tuning, as it generates relatively high-quality text. ChatGLM2 (Zeng et al., 2022) obtains a GFI value of 1.4411 under the 1-Shot setting, slightly lower than our model. This indicates that ChatGLM2 exhibits considerable text generation capabilities even with only a small amount of fine-tuning samples. Bloomz (Muennighoff et al., 2022) exhibits relatively average performance. It achieves GFI values of 2.9808 and 7.1954 under the zero-shot and 1-shot settings, respectively, significantly higher than other LLMs. The higher GFI values suggest that the text generated by Bloomz is more complex and less understandable, possibly due to performance degradation caused by its cross-lingual generalization.

Finally, we conducted an analysis of the results obtained by PowerPulse and the state-of-the-art model, ChatGPT, in the text generation task. PowerPulse achieved a GFI value of 1.1410 under zero-shot conditions, signifying its excellent performance in maintaining text readability. In contrast, ChatGPT yielded a GFI of 1.4456, indicating slightly lower readability in zero-shot text generation compared to PowerPulse. This suggests that PowerPulse demonstrates robustness in generating coherent and contextually appropriate responses without the need for any task-specific prompts. While ChatGPT performed admirably, it falls marginally behind PowerPulse in this particular scenario. PowerPulse maintains a competitive GFI value of 1.2380, showcasing its adaptability to 1-shot text generation tasks. Meanwhile, ChatGPT performs strongly with a GFI of 1.3955, albeit slightly lower than PowerPulse. In the context of 1-shot text generation, both models demonstrate robust performance, emphasizing PowerPulse's effectiveness in rapidly adapting to limited examples, rendering it valuable for scenarios requiring swift adjustment.

In summary, PowerPulse demonstrates outstanding performance in the text generation task of the power energy domain, with much lower GFI values compared to the other Open source LLMs, including LLaMA (Touvron et al., 2023) and ChatGLM2 (Zeng et al., 2022). This suggests that our improved approach based on LLaMA, that is, fine-tuning the large model based on the domain dataset and the domain instruction set, can effectively enhance the domain adaptability of LLMs to better handle the domain text generation task. When the quality of the constructed dataset is high enough, this approach can be equally well applied to the construction of large models for other specific domains. Meanwhile, Bloomz (Muennighoff et al., 2022) shows relatively average performance in this task and may require further improvements in its cross-lingual generalization capabilities.

Table 4 displays the ROUGE values of PowerPulse and other state-of-the-art LLMs in the summary extraction task, with experiments conducted under both the zero-shot and 1-shot settings. The ROUGE -1, ROUGE -2, and ROUGE -L scores represent the ROUGE -N values for 1-gram, 2-gram, and L-gram cases, respectively. The best results indicated in bold and the second-best results are underlined.

Based on the findings from Table 4, the following conclusions can be drawn.

By observing the table, it is evident that our PowerPulse achieves significantly higher ROUGE -1, ROUGE -2, and ROUGE -L scores (in bold) compared to other advanced LLMs under both zero-shot and 1-shot settings. Especially in terms of ROUGE -1 and ROUGE -L evaluation metrics,

PowerPulse attains values of 0.3945 and 0.3539 (zero-shot), as well as 0.2902 and 0.2373 (1-shot), far surpassing the performance of other models. This indicates that the summaries generated by PowerPulse in the summary extraction task are more akin to the reference summaries, exhibiting higher similarity and coverage, thus showcasing its superiority in summary generation.

LLaMA's performance in the summary extraction task is relatively inferior compared to other advanced models, with ROUGE -1 and ROUGE -L values both below 0.2, indicating insufficient similarity and coverage with the reference summaries. ChatGLM2 slightly outperforms LLaMA in ROUGE -1 and ROUGE -L, yet it still falls far short of PowerPulse. Although its ROUGE -1 score is relatively high under the 1-shot setting, all its ROUGE scores remain significantly lower than those of PowerPulse. Bloomz (Muennighoff et al., 2022) performs poorly in the summary extraction task. As evident from the table, Bloomz's ROUGE scores under both zero-shot and 1-shot settings are significantly lower than those of other LLMs. This suggests that the summaries generated by Bloomz exhibit lower similarity with the reference summaries, indicating inferior summary quality compared to other models.

Finally, we conducted an analysis of the results obtained by PowerPulse and the state-of-the-art model, ChatGPT. In the zero-shot setting, upon evaluating the ROUGE metrics for summary extraction, it becomes evident that PowerPulse surpasses ChatGPT across multiple ROUGE scores. Specifically, PowerPulse achieves notably higher scores in ROUGE-1 and ROUGE-L when compared to ChatGPT. This indicates that PowerPulse excels in capturing essential content and maintaining the coherence of automatically generated summaries, underscoring its effectiveness in zero-shot summary extraction tasks. While ChatGPT exhibits commendable performance, particularly in ROUGE-2 scores. In the one-shot setting, ChatGPT's ROUGE scores are indeed impressive, but PowerPulse maintains a consistent competitive edge. In fact, PowerPulse either matches or surpasses ChatGPT's performance in these scenarios. This highlights the robustness and adaptability of PowerPulse in producing high-quality summaries, even when provided with limited examples.

In conclusion, PowerPulse demonstrates notable superiority in the summary extraction task of the power energy domain, generating summaries that closely resemble the reference summaries with higher similarity and coverage. From the calculation, Rouge-1 can represent the informativeness of the generated summaries and Rouge-L represents the fluency of the summaries. As can be seen from Table 4, the summaries generated by PowerPulse outperform the other open source LLMs both in terms of informativeness and fluency, which suggests that both incremental pre-training based on the power energy domain dataset and fine-tuning of instructions based on the domain greatly enriched the model's amount of knowledge about the power energy domain, which further improves the model's performance in the task of summary extraction. Thinking differently, acting in the same way in different specialized domains, it can be envisioned that our approach can likewise be equally effective in different domains. This fully demonstrates the value and impact of our research on a wide range of applications in different domains. Positioning it as a compelling alternative to ChatGPT for applications that demand top-quality, contextually relevant summaries.

Table 5 presents the Accuracy values of our PowerPulse and other state-of-the-art LLMs on the topic classification task. Where the best results are in bold and the second best results are underlined. By comparing the data in the table, it is evident that PowerPulse achieves significantly higher performance (in bold) compared to other LLMs. Specifically, PowerPulse attains accuracy rates of 23.76% in the zero-shot scenario and 33.64% in the 1-shot scenario, outperforming other LLMs by a wide margin.

LLaMA demonstrates relatively lower performance in this task, with accuracy rates of only 14.85% in the zero-shot scenario and 19.84% in the 1-shot scenario. While ChatGLM2 slightly surpasses LLaMA, it still lags behind PowerPulse, achieving accuracy rates of 20.79% and 23.76% in the zero-shot and 1-shot scenarios, respectively. Although Bloomz shows improvement in the 1-shot scenario, it still falls short of the accuracy levels of PowerPulse, with zero-shot and 1-shot accuracies of 5.94% and 12.87%, respectively. These results clearly demonstrate the outstanding performance of PowerPulse in the topic classification task, significantly outshining other LLMs. The superior performance of PowerPulse can be attributed to its fine-tuning specifically on the field of power energy, leveraging relevant pre-training data and domain-specific fine-tuning dataset, which enhance its performance in classifying topics related to power energy industry.

In the context of topic classification, both zero-shot and one-shot settings provide valuable insights into the capabilities of PowerPulse and ChatGPT. In the zero-shot scenario, ChatGPT exhibits a slightly higher accuracy rate compared to PowerPulse. ChatGPT achieves an accuracy rate of 25.59%, while PowerPulse closely follows with an accuracy rate of 23.76%. This suggests that ChatGPT may have an advantage in classifying text accurately when no task-specific examples are provided. It showcases ChatGPT's ability to generalize well to new tasks, drawing upon its extensive pre-trained knowledge. However, it is noteworthy that PowerPulse remains competitive in this setting, showcasing its capacity to adapt

TABLE 5 Accuracy values of LLMs in topic classification task under zero and 1-shot settings.

Models	Zero-shot (%)	1-shot (%)
LLaMA (Touvron et al., 2023)	14.85	19.84
ChatGLM2 (Du et al., 2021)	20.79	23.76
Bloomz (Muennighoff et al., 2022)	5.94	12.87
ChatGPT (Ouyang et al., 2022)	25.59	<u>32.67</u>
PowerPulse (Ours)	<u>23.76</u>	33.64

to unfamiliar classification tasks without the need for extensive task-specific training data. In the one-shot scenario, PowerPulse excels and outperforms ChatGPT. PowerPulse achieves a remarkable accuracy rate of 33.64%, while ChatGPT achieves an accuracy rate of 32.67%. This signifies PowerPulse's robustness and adaptability when given just a single example for a classification task. Its ability to provide accurate classifications with limited examples is a notable advantage, particularly in scenarios where data availability is limited or where quick adaptation to new tasks is required.

In conclusion, the exceptional performance of PowerPulse in the topic classification task is attributed to its specialized training in the field power energy, resulting in a distinct advantage in identifying and classifying unfamiliar topics. On the other hand, other LLMs' performance suffers due to the lack of domain-specific training data, leading to suboptimal performance in domain-specific applications. This underscores the significance of PowerPulse as a valuable reference and model for the development of domain-specific chat models. Overall, the comparison between PowerPulse and ChatGPT in text classification tasks reveals a nuanced picture. ChatGPT displays its strength in zero-shot scenarios, showcasing its generalization abilities. On the other hand, PowerPulse shines in the one-shot scenario, demonstrating its adaptability and accuracy with minimal examples. Therefore, the choice between PowerPulse and ChatGPT for text classification tasks may depend on the specific requirements of the application, the availability of training data, and the desired balance between generalization and adaptability.

4.4 | Case studies

To better illustrate the performance of PowerPulse in various tasks, we conducted case studies comparing it with other LLMs. The experimental results are summarized in Tables 6 and 7.

From the case studies presented in Table 6, it is evident that PowerPulse demonstrates remarkable performance in the task of text generation. Our model displays a strong familiarity with domain-specific terminology, adapts well to the language style, and exhibits a high level of professionalism. In contrast, ChatGLM lacks a clear language style, resulting in overly general contents that lack specificity. LLaMA slightly deviates from the main theme during text generation and lacks domain-specific targeting, which renders it inadequate for meeting the requirements of specific domains. Furthermore, ChatGPT's performance also reveals a lack of a distinct language style, potentially resulting in the production of content that lacks the necessary granularity. This limitation poses challenges when tackling text generation tasks in specific domains, where the use of domain-specific terminology and style is essential. Therefore, the key advantage of PowerPulse in text generation lies in its precise use of power energy domain terminologies and language styles, enabling it to effectively fulfil the requirements of text generation tasks.

Based on Table 7, it is evident that PowerPulse excels in the task of summary extraction, successfully capturing and presenting key information from the given data. In comparison, LLaMA and ChatGLM demonstrate weaker performance on this task, as they fail to accurately extract the summary information. LLaMA mentions the time frame of Poland's nuclear energy policy advancement and the deployment of six reactors but does not include information about the environmental impact assessment report for the nuclear power plant. While ChatGLM mentions the submission of the environmental impact assessment report by the PEJ company, but the information provided is not sufficiently concise and succinct. Regarding ChatGPT's performance in this task, its primary strength lies in offering a general overview of the information. However, relative to the reference summary, ChatGPT may lack some specific details. For tasks in specialized domains, further post-editing or adjustments may be required to ensure the accuracy and comprehensiveness of the summary.

As a result, PowerPulse's outstanding performance in summary extraction makes it a highly promising domain-specific question-answering model, particularly when dealing with information related to the power energy sector. Its advantages in this regard are particularly pronounced.

5 | CONCLUSIONS AND FUTURE WORK

5.1 | Conclusions

In this paper, we present a novel power energy chat model named PowerPulse, built upon the LLaMA architecture and fine-tuned with Chinese power sector-specific knowledge. Through comprehensive experiments, we demonstrate PowerPulse's remarkable performance in tasks such as text generation, summary extraction, and topic classification. However, we acknowledge certain limitations of PowerPulse, such as potential vulnerability to updates and changes in domain-specific knowledge. Given the evolving nature of policies and technologies in the energy and power sectors, continuous updates to the model's fine-tuning data and knowledge base are essential to ensure its ongoing effectiveness.

5.2 | Future work

Furthermore, while our evaluation was thorough, we observed that PowerPulse's ability to handle multi-turn conversations was not fully showcased, primarily due to the scarcity of multi-turn dialogue datasets. The limited availability of such data may affect PowerPulse's performance in

TABLE 6 Case study results of various LLMs for text generation task.

Models	Instances (in Chinese)	Instances (translated to English)
Input	请以给定的题目，写一篇内容400字左右的精彩报道。题目为:<碳中和下氢能行业深度报告>	Write a great report with a content of about 400 words on the given topic. The title is: <In-depth report on the hydrogen energy industry under carbon neutrality>
LLaMA (Touvron et al., 2023)	在当前全球气候变化日益严峻的情况下，低碳环保已经成为了人们关注的重要话题之一。而作为一种清洁、高效的新能源技术——氢能在近年来得到了广泛的应用和发展。据最新数据显示，目前我国已经建成了一批规模化生产氢气的企业，并且正在积极推进相关政策的支持与推广工作。未来几年内，随着新能源产业的发展以及政府对绿色发展的支持力度不断加大，氢能将成为一个重要的发展方向。	Under the current situation of increasingly severe global climate change, low-carbon environmental protection has become one of the important topics of concern. As a clean and efficient new energy technology, hydrogen energy has been widely used and developed in recent years. According to the latest data, China has built a number of large-scale productions of hydrogen enterprises, and is actively promoting the support and promotion of related policies. In the next few years, with the development of new energy industry and increasing government support for green development, hydrogen energy will become an important development direction.
ChatGLM (Du et al., 2021)	氢能行业深度报告近年来，随着环保政策的不断推动和消费者对绿色能源的需求不断增加，氢能行业逐渐兴起。作为一种清洁、高效、可持续的能源形式，氢能技术在各个领域得到了广泛的应用，如汽车、能源、食品等。本报告将深度分析碳中和下氢能行业的发展趋势和挑战，为您带来一份精彩报告 ...	Hydrogen Energy Industry In-depth Report In recent years, the hydrogen energy industry has gradually emerged with the continuous promotion of environmental protection policies and increasing consumer demand for green energy. As a clean, efficient and sustainable form of energy, hydrogen energy technology has been widely used in various fields, such as automotive, energy and food. This report will deeply analyze the development trends and challenges of the hydrogen energy industry under carbon neutrality and bring you an exciting report ...
ChatGPT (Ouyang et al., 2022)	随着全球应对气候变化和向可持续能源转型的迫切需要，氢能产业已成为实现碳中和的有力参与者。本报告深入研究了该行业的现状，探讨了其发展前景，并揭示了其在实现全球碳中和目标中的关键作用。 碳中和，即通过清除或抵消来平衡碳排放的概念，已成为全球应对气候变化的首要目标。在此背景下，经常被称为“未来燃料”的氢能作为碳中和等式的重要组成部分，正日益受到重视...	As the world grapples with the pressing need to combat climate change and transition towards sustainable energy sources, the hydrogen energy industry has emerged as a powerful player in the quest for carbon neutrality. This report delves into the industry's current state, explores its growth prospects, and unveils its pivotal role in achieving global carbon neutrality goals. Carbon neutrality, the concept of balancing carbon emissions with removal or offsetting, has become a paramount global objective in the battle against climate change. In this context, hydrogen energy, often referred to as the “fuel of the future,” is gaining prominence as a vital component of the carbon-neutral equation ...
PowerPulse (Ours)	近日，由中国化工与能源研究会、国家可再生能源中心联合主办的“2023年全球氢能产业发展研讨会在京召开”。会上发布了《2020-2025年世界氢能产业市场前景分析及投资机会研究报告》(以下简称《报告》)。《报告》指出，在实现二氧化碳排放达峰后，到2030年前将达到零增长的目标之后，人类社会进入低碳化时代，并逐步转向净零碳经济体系。随着气候变化问题的日益凸显以及可再生资源逐渐成为主要能源来源的趋势，氢能在未来将成为重要的清洁替代燃料之一...	Recently, the “2023 Global Hydrogen Energy Industry Development Seminar” was held in Beijing, which was jointly organized by China Chemical Industry and Energy Research Society and National Renewable Energy Center. The “2020-2025 World Hydrogen Energy Industry Market Outlook Analysis and Investment Opportunities Research Report” (hereinafter referred to as the “Report”) was released at the meeting. The Report pointed out that after achieving the goal of zero growth by 2030 following the peaking of carbon dioxide emissions, human society has entered the era of decarbonization and is gradually shifting to a net-zero carbon economic system. With the increasing prominence of climate change issues and the trend of renewable resources gradually becoming the main source of energy, hydrogen energy will become one of the important clean alternative fuels in the future ...

handling complex multi-turn conversations. In practical applications, multi-turn dialogues are crucial for domain question-answering systems, as users often require continuous interactions and in-depth discussions. To address these challenges, we plan to further enhance PowerPulse by incorporating more domain-specific corpora and diversifying the instruction fine-tuning datasets. Additionally, we aim to assess the model's

TABLE 7 Case study results of various LLMs for summary extraction task.

Models	Instances (in Chinese)	Instances (translated to English)
Input	据世界核新闻网站2022年3月31日报道，波兰国有企业PEJ公司正在推进该国2040年前部署6座反应堆的能源政策。2022年3月29日，PEJ公司向波兰环境保护总局(GDO)提交了波兰首座核电厂环境影响评价(EIA)报告。环境保护总局表示，环评报告是环境影响评估程序中的关键文件。该报告分析了在波美拉尼亚省霍切沃市、格涅维诺市和科尔科瓦市建设和运营波兰首座核电厂的环境影响。该核电厂总装机容量为375万千瓦。2021年12月，霍切沃市卢比亚托沃镇和科帕利诺镇被定为波兰首座大型核电厂的首选厂址。 参考摘要:波兰提交首座核电厂环评报告。	Poland's state-owned PEJ is moving ahead with its energy policy of deploying six reactors by 2040, World Nuclear News reported on March 31, 2022. On March 29, 2022, PEJ submitted to the Polish General Directorate for Environmental Protection (GDO) the Environmental Impact Assessment (EIA) for Poland's first nuclear power plant. The General Directorate for Environmental Protection said that the EIA report is a key document in the environmental impact assessment procedure. The report analyzes the environmental impact of the construction and operation of Poland's first nuclear power plant in the municipalities of Hoczewo, Gniewno and Kolkowa in Pomeranian Voivodeship. The total installed capacity of the plant will be 3.75 million kW. In December 2021, the towns of Lubiatowo and Kopalino in the municipality of Hoczewo were designated as the preferred sites for Poland's first large-scale nuclear power plant. Reference Abstract: Poland submits EIA report for first nuclear power plant.
LLaMA (Touvron et al., 2023)	波兰 PEJ 公司推进核能政策，到 2048 年部署6个反应堆。	Poland's PEJ Company Advances Nuclear Energy Policy with 6 Reactor Deployment by 2048.
ChatGLM (Du et al., 2021)	PEJ公司推进波兰2040年前部署6座反应堆的能源政策，提交波兰首座核电厂环境影响评价报告。	PEJ advances Poland's energy policy of deploying six reactors by 2040, submits EIA report for Poland's first nuclear power plant.
ChatGPT (Ouyang et al., 2022)	波兰计划2040年前建设6座反应堆的核电厂，首座选址进展顺利。	Poland plans to build a six-reactor nuclear power plant by 2040, with the first site well underway.
PowerPulse (Ours)	波兰PEJ公司提交首个核电站环境影响研究报告。	Poland's PEJ submits its first environmental impact study for a nuclear power plant.

performance on more domain tasks, based on both our model-centric evaluation and human preference-based evaluation. By continuously refining PowerPulse, we believe it will play a pivotal role in the field of power energy, meeting the demands of various tasks in this specialized field.

While PowerPulse is specifically designed for the power energy domain and holds significant potential in enhancing the professionalism of power and energy policy consulting and decision-making, its fundamental principles and capabilities can be extended to numerous other specialized domains. Although ChatGPT cannot be utilized offline, the availability of various feasible LLMs within the open-source community (Wang et al., 2023), such as LLaMa (Touvron et al., 2023), ChatGLM (Du et al., 2021; Zeng et al., 2022), and Alpaca (Taori et al., 2023), with relatively low training costs and reduced hardware requirements, makes it viable to train custom domain-specific LLMs. Specialized language models like PowerPulse can provide highly accurate, context-aware, and domain-specific language generation capabilities, potentially ushering in transformative changes across various industries.

Furthermore, substantial achievements have already been made by fine-tuning large models from the aforementioned open-source resources in fields such as healthcare, finance, legal, scientific research, and coding domains. Some potential broader applications and impacts encompass.

Medical and healthcare (Li et al., 2023a; Singhal et al., 2023; Wang et al., 2023; Xiong et al., 2023; Yunxiang et al., 2023): Specialized models can assist healthcare professionals in generating accurate medical reports, interpreting patient data, and providing domain-specific information to patients.

Legal (Gentile, 2023; Huang et al., 2023): In the legal sector, such models aid in drafting legal documents, contracts, and summarizing complex legal cases.

Finance (Wu et al., 2023; Xie et al., 2023; Yang et al., 2023): Specialized language models can assist financial analysts in generating financial reports, market analyses, and investment recommendations.

Scientific research (Taylor et al., 2022): Researchers can benefit from these models for generating research papers, summarizing scientific findings, and automating data analysis.

Education: Specialized models can support educators by generating course materials, providing explanations for complex subjects, and assisting in the creation of educational content.

These domain-specific LLMs serve as inspirations for more extensive models across various domains, paving the way for broader applications and impact.

ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Project of Yunnan Power Grid Co., Ltd. under grant no. YNKJXM20222254; the Science Foundation of Young and Middle-aged Academic and Technical Leaders of Yunnan under grant no. 202205AC160040; the Science Foundation of Yunnan Jinzhi Expert Workstation under grant no. 202205AF150006; the Major Project of Yunnan Natural Science Foundation under grant no. 202302AE09002003; the Open Foundation of Yunnan Key Laboratory of Software Engineering under grant no. 2023SE101; the Science Foundation of 'Knowledge-Driven Intelligent Software Engineering Innovation Team'.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Xuan Zhang  <https://orcid.org/0000-0003-2929-2126>

REFERENCES

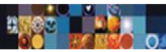
- Aghajanyan, A., Zettlemoyer, L., & Gupta, S. (2020). Intrinsic dimensionality explains the effectiveness of language model fine-tuning. arXiv preprint arXiv:2012.13255.
- Ali, N., Tubaishat, A., Al-Obeidat, F., Shabaz, M., Waqas, M., Halim, Z., Rida, I., & Anwar, S. (2023). Towards enhanced identification of emotion from resource-constrained language through a novel multilingual BERT approach. *ACM Transactions on Asian and Low-Resource Language Information Processing*, All Works, 5790.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
- Burke, L., Pazdernik, K., Fortin, D., Wilson, B., Goychayev, R., & Mattingly, J. (2021). NukeLM: Pre-Trained and Fine-Tuned Language Models for the Nuclear and Energy Domains. arXiv preprint arXiv:2105.12192.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.
- Cui, Y., Yang, Z., & Yao, X. (2023). Efficient and effective text encoding for Chinese LLaMA and alpaca. arXiv preprint arXiv:2304.08177.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M., & Zhou, B. (2023). Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. arXiv preprint arXiv:2305.14233.
- Du, K., Zhang, X., Gao, C., Zhu, R., Nong, Q., Yang, X., & Yin, C. (2023). GIMM: A graph convolutional network-based paraphrase identification model to detecting duplicate questions in QA communities. *Multimedia Tools and Applications*, 1–28.
- Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2021). Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360.
- DuBay, W. H. (2004). Judges scold lawyers for bad writing. *Plain Language At Work Newsletter (Impact Information)*(8).
- Gentile, G. (2023). *LawGPT? How AI is reshaping the legal profession*. Impact of Social Sciences Blog. <https://blogs.lse.ac.uk/impactofsocialsciences/2023/06/08/lawgpt-how-ai-is-reshaping-the-legal-profession/>
- Gupta, T., Zaki, M., Krishnan, N. A., & Mausam. (2022). MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1), 102.
- Halim, Z., Waqar, M., & Tahir, M. (2020). A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowledge-Based Systems*, 208, 106443.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790–2799). PMLR.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.0968.
- Huang, Q., Tao, M., An, Z., Zhang, C., Jiang, C., Chen, Z., Wu, Z., & Feng, Y. (2023). Lawyer LLaMA Technical Report. arXiv preprint arXiv:2305.15062.
- Jain, A., Meenachi, D. N., & Venkatraman, D. B. (2020). NukeBERT: A pre-trained language model for low resource nuclear domain. arXiv preprint arXiv:2003.13821.
- Ji, Y., Deng, Y., Gong, Y., Peng, Y., Niu, Q., Zhang, L., Ma, B., & Li, X. (2023). Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. arXiv preprint arXiv:2303.14742.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
- Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890.
- Li, J., Wang, X., Wu, X., Zhang, Z., Xu, X., Fu, J., Tiwari, P., Wan, X., & Wang, B. (2023). Huatuo-26M, a Large-scale Chinese Medical QA Dataset. arXiv preprint arXiv:2305.01526.
- Li, W., Wu, W., Chen, M., Liu, J., Xiao, X., & Wu, H. (2022). Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. arXiv preprint arXiv:2203.05227.

- Li, X. L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 4582–4597).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pre-training approach. arXiv preprint arXiv:1907.11692.
- Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1), 8–30.
- Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Alnubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2022). Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786.
- OpenAI. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.
- Rahman, A. U., & Halim, Z. (2023). Identifying dominant emotional state using handwriting and drawing samples by fusing features. *Applied Intelligence*, 53(3), 2798–2814.
- Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczelcha, E., Kim, T., Chhablani, G., Nayak, N., ... Rush, A. M. (2021). Multitask Prompted Training Enables Zero-Shot Task Generalization. arXiv preprint arXiv:2110.08207.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Blaise Aguera, Y., ... Natarajan, V. (2023). Towards Expert-Level Medical Question Answering With Large Language Models. arXiv preprint arXiv:2305.09617.
- Tahir, M., Halim, Z., Waqas, M., & Tu, S. (2023). On the effect of emotion identification from limited translated text samples using computational intelligence. *International Journal of Computational Intelligence Systems*, 16(1), 107.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., & Hashimoto, T. B. (2023). *Stanford alpaca: An instruction-following llama mode*.
- Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., & Stojnic, R. (2022). Galactica: A large language model for science. arXiv preprint arXiv:2211.09085.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Edouard Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010).
- Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., & Liu, T. (2023). Huatuo: Tuning LLaMA model with Chinese medical knowledge. arXiv preprint arXiv:2304.06975.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned Language Models are Zero-Shot Learners. arXiv preprint arXiv:2109.01652.
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.
- Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. arXiv preprint arXiv:2306.05443.
- Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., & Shen, D. (2023). Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097.
- Xu, M. (2023). textgen: Implementation of language model finetune. <https://github.com/shibing624/textgen>
- Yang, H., Liu, X. Y., & Wang, C. D. (2023). FinGPT: Open-Source Financial Large Language Models. arXiv preprint arXiv:2306.06031.
- Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive Mirage: A Review of Hallucinations In Large Language Models. arXiv preprint arXiv:2309.06794.
- Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., & You, Z. (2023). Chatdoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge. arXiv preprint arXiv:2303.14070.
- Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Zhang, P., Dong, Y., & Tang, J. (2022). Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414.
- Zhang, G., Shi, Y., Liu, R., Yuan, R., Li, Y., Dong, S., Shu, Y., Li, Z., Wang, Z., Lin, C., Huang, W., & Fu, J. (2023). Chinese open instruction generalist: A preliminary release. arXiv preprint arXiv:2304.07987.

AUTHOR BIOGRAPHIES

ChunLin Yin received his Master's degree in Computer Science from Yunnan University. He is an engineer at the Electric Power Research Institute of Yunnan Power Grid Co., Ltd. He has published more than 20 papers and has been granted more than 10 invention patents. His main research methods are Natural Language Processing and Continual Learning.

KunPeng Du received the M.S. degree in electronic information from Yunnan University. He is currently a doctoral student at the School of Software at Yunnan University. His research interests include LLMs Autonomous Agents and spoken language understand.



Qiong Nong received the B.S. degree in software engineering from Guilin University of Technology and is currently studying at Yunnan University, a graduate student majoring in software engineering at the School of Software. Her research interest is natural language processing (NLP).

Hongcheng Zhang (1989), male, originally from Gong'an, Hubei, holds the title of Engineer and has a Master's degree. His main research direction is policy research and management innovation.

Li Yang received the MA. Eng degree in computer science from Kunming University of Science and Technology in 2011, Kunming, China. She is currently work for the Electric Power Research Institute of Yunnan Power Grid as an senior engineer. She has been a principal investigator for more than 10 scientific research projects. She is the author of more than 30 articles. Her main research interests are digital transformation.

Bing Yan (1987), male, originally from Yiliang, Yunnan, holds the title of Senior Engineer and has a Master's degree. His main research direction is policy research and management innovation.

Xiang Huang received the M.S. degree in software engineering from Yunnan University, Kunming, China. He is now working at the Electric Power Research Institute of Yunnan Power Grid Co., Ltd. His research interests include reinforcement learning and software development.

Xiaobo Wang received the B.S. degree in software engineering from Hunan University of Arts and Sciences and is currently pursuing a Master's degree in Software Engineering at the School of Software, Yunnan University. His research interests are natural language processing (NLP) and knowledge graphs (KG).

Xuan Zhang received the B.S. and M.S. degrees in computer science and the Ph.D. degree in system analysis and integration from Yunnan University, Kunming, China. She is currently a Professor with the School of Software, Yunnan University, Yunnan, China. She is also the Core Scientist of the Yunnan Key Laboratory of Software Engineering and the Yunnan Software Engineering Academic Team. She has been a principal investigator for more than 30 national, provincial, and private grants and contracts. She is the author of three books and more than 100 articles. Her research interests include knowledge graph (KG), natural language processing (NLP), and recommendation systems.

How to cite this article: Yin, C., Du, K., Nong, Q., Zhang, H., Yang, L., Yan, B., Huang, X., Wang, X., & Zhang, X. (2023). PowerPulse: Power energy chat model with LLaMA model fine-tuned on Chinese and power sector domain knowledge. *Expert Systems*, e13513. <https://doi.org/10.1111/exsy.13513>