

ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification

Biao Zhao^{a,1}, Weiqiang Jin^{a,1}, Javier Del Ser^b, Guang Yang^{c,d,e,*}

^a School of Information and Communications Engineering, Xi'an Jiaotong University, Innovation Harbour, Xi'an, 710049, Shaanxi, China

^b TECNALLIA, Basque Research & Technology Alliance (BRTA), Derio, 48160, Spain

^c Bioengineering, Imperial College London, London, SW7 2BX, UK

^d Imperial-X, Imperial College London, London, W12 7SL, UK

^e National Heart and Lung Institute, Imperial College London, London, SW3 6LY, UK

ARTICLE INFO

Communicated by N. Zeng

Keywords:

Agricultural text classification
Very large pre-trained language model
Generative Pre-trained Transformer (GPT)
ChatGPT
GPT-4

ABSTRACT

In the era of sustainable smart agriculture, a vast amount of agricultural news text is posted online, accumulating significant agricultural knowledge. To efficiently access this knowledge, effective text classification techniques are urgently needed. Deep learning approaches, such as fine-tuning strategies on pre-trained language models (PLMs), have shown remarkable performance gains. Nonetheless, these methods face several complex challenges, including limited agricultural training data, poor domain transferability (especially across languages), and complex and expensive deployment of large models. Inspired by the success of recent ChatGPT models (e.g., GPT-3.5, GPT-4), this work explores the potential of applying ChatGPT in the field of agricultural informatization. Various crucial factors, such as prompt construction, answer parsing, and different ChatGPT variants, are thoroughly investigated to maximize its capabilities. A preliminary comparative study is conducted, comparing ChatGPT with PLMs-based fine-tuning methods and PLMs-based prompt-tuning methods. Empirical results demonstrate that ChatGPT effectively addresses the mentioned research challenges and bottlenecks, making it an ideal solution for agricultural text classification. Moreover, ChatGPT achieves comparable performance to existing PLM-based fine-tuning methods, even without fine-tuning on agricultural data samples. We hope this preliminary study could inspire the emergence of a general-purpose AI paradigm for agricultural text processing.

1. Introduction

With the rapid development of sustainable smart agriculture ecosystem, the quantity of various news contents related to agricultural themes on the Internet has undergone an explosive increase. Such a vast quality of unstructured data contains already latent historical knowledge, helping us precisely study natural hazards and mitigate potential agricultural risks. Artificial intelligence-based agricultural text classification enables managing these massive Internet agricultural news automatically and makes these massive unstructured data easily indexable, which is a crucial step for agricultural digitization and agricultural Internet of Things [1,2].

In recent years, mainstream agricultural document processing techniques have relied on deep representation learning methods, particularly pre-trained language models (PLMs) like BERT, BART, and T5 [3–5]. Xu et al. [6] introduced the AETS-LSTM model for predicting

agricultural export fluctuations. Cao et al. [7] utilized BERT with a symmetrical structure to analyze consumer sentiment towards agricultural products. Leong et al. [8] used the CRAFT model for extracting essential information from agricultural regulatory documents [9,10]. Jiang et al. [1] proposed a BERT-based text classification network for indexing French bulletins. Deep representation learning methods hold promise for various agricultural informatics applications.

Unfortunately, PLMs-based fine-tuning solutions encounter challenging issues in practical model development and deployment [11,12]. Insufficient and poor quality training data can reduce model performance, while obtaining high-quality annotated data is time-consuming and labor-intensive. Even with sufficient training data, supervised learning models have limited generalization capabilities and lack scenario transferability, especially cross-linguistic capacity. Additionally, PLMs' large parameter volumes make deployment complex and power-intensive, requiring high-performance equipment like GPUs and TPUs.

* Corresponding author.

E-mail addresses: biaozhao@xjtu.edu.cn (B. Zhao), weiqiangjin@stu.xjtu.edu.cn (W. Jin), javier.delser@tecnalia.com (J. Del Ser), g.yang@imperial.ac.uk (G. Yang).

¹ Both the first two authors, Biao Zhao and Weiqiang Jin, made equal contributions to this work.

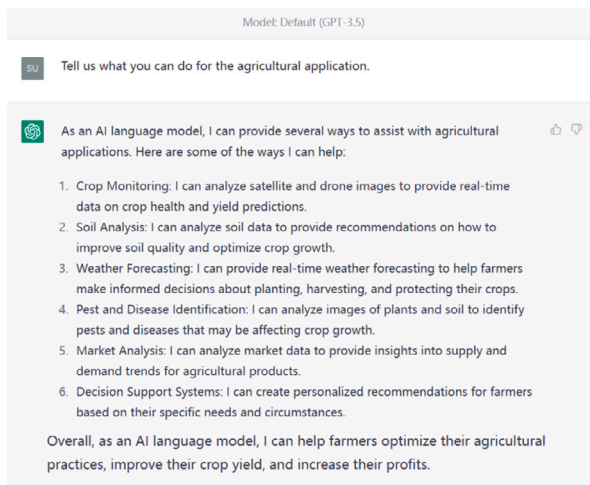


Fig. 1. Valuable suggestions advised by ChatGPT for assisting farmers and market regulator in better governing agricultural affairs (Query Date: 2023.3.16).

For example, the largest T5 model has over 11 billion parameters, which is 100 times the number of parameters of the BERT-base model. These limitations highlight the shortcomings of mainstream PLMs-based agricultural text classification methods in achieving General Purpose Artificial Intelligence (GPAI) standards.

The limitations of existing agricultural document processing techniques hinder their effectiveness in various application scenarios, particularly in agricultural text classification. Recently, OpenAI's ChatGPT-family chatbots have revolutionized the academic community, especially in natural language processing (NLP) tasks. ChatGPT,² a powerful language model based on the Transformer architecture, has garnered attention for its comprehensive responses to complex questions and its abilities in multilingual translation, poetry generation, and code generation [13,14]. Its language comprehension and generation capabilities have gained popularity in cross-disciplinary research, such as radiology diagnosis [15] and sentiment analysis of surgical diseases [16].

Considering ChatGPT's capabilities, it is natural to explore its potential in optimizing sustainable agricultural applications. According to GPT-3.5 (a standard model in ChatGPT-family), it can contribute to weather forecasting, pest and disease identification, and market analysis in agriculture (among other tasks) (Fig. 1). With the potential applications of ChatGPT in smart agriculture, there is a need for principled explorations to assess its contribution to sustainable agricultural practices. Therefore, we aim to investigate ChatGPT's potential in concise agricultural text classification in this study.

By doing so, our experiments mainly investigate the potential power of ChatGPT (i.e. GPT-3.5 by default) [17] and its extension (i.e. GPT-4) [14] for classifying the agricultural-related documents. Notably, along with the proposed ChatAgri, this paper also provides a brand-new paradigm which is distinguished from existing methods. Through a series comparative experiments of ChatAgri with a range of mainstream text classification models, including classic fine-tuned PLMs [18,19] and prompt-learning based on auto-regressive generative PLMs [20–22], we systematically evaluated and investigated the superiority of ChatGPT in agricultural text classification tasks, which distinguished it significantly from other methods.

Furthermore, we have investigated extensive literature related to ChatGPT-based question answering (QA) [23–26] and the prompt learning scheme [20,27,28], and arrived at the following conclusions: Most language understanding tasks based on ChatGPT can be

categorized as a new form of Prompt Learning based on PLMs. Specifically, regarding the adopted ChatGPT interface as a parameters-frozen large-scale PLM, the overall procedure are extremely similar to the prompt-tuning paradigm described in the survey of Liu et al. [20]. Fig. 2 gives a clear illustration of the major similarities and distinguishes between ChatGPT-based NLP paradigm, (a) and MLM prompt-tuning paradigm, (b), through a typical example of the agricultural food comment sentiment analysis task. As depicted in part. (c) of Fig. 2, the MLM prompt-tuning paradigm can be divided into three primary procedures: template engineering, pre-trained language models reasoning, and answer mapping engineering [20]. As shown in part. (b) of Fig. 2, the general NLP research related to ChatGPT can be organized into the following several phases in our experiments [15,29]: (1) prompting question construction engineering; (2) ChatGPT Q&A inference; (3) answer normalization engineering (alias. answer alignment). Thus, several core factors were considered to be optimized:

- (1). Due to that interacting with ChatGPT involves providing instructions through human response, based on previous ChatGPT prompting works [24,25,30], we have designed several appropriate task-specific inquiries to intuitively trigger the understanding capability of ChatGPT;
- (2). As the textual generations of ChatGPT are essentially human-like natural language, they differ greatly when it comes to specific tasks. So, a accurate label mapping strategy from ChatGPT outputs to the final classified categories are needed to be developed. In our experiments, we devised two novel answer mapping strategies for this critical step for the answer alignment engineering.

To evaluate extensive data in various agricultural sub-fields, sourced mainly comes from Internet news covering topics ranging from insect pests, and natural hazards to agricultural market comments. Further, even in cases multi-language corpora are tested, experiments validate that the proposed ChatAgri still features a significant transferring effectiveness in cross-linguistic scenarios.

In summary, our experiments provide a preliminary study of ChatGPT on agricultural text classification to gain a better understanding of it, and reported a systematic analysis according to the corresponding empirical results. We believe that by exploring how ChatGPT can contribute to agricultural production and management through text classification tasks such as pest and disease identification, agricultural news categorization, and market comment analysis, we can demonstrate the feasibility of ChatGPT in advancing agricultural practices, thereby paving the way for a more efficient and sustainable smart agriculture.

The novel ingredients of this work can be summarized as follows:

- Motivated by the progress of large pre-trained language models like ChatGPT, we conducted a preliminary study on agricultural text classification and proposed ChatAgri as a ChatGPT-based solution.
- ChatAgri achieved competitive performance compared to existing approaches, demonstrating impressive semantic understanding and intelligent reasoning in specific cases.
- Zero-shot learning experiments highlighted ChatAgri's potential, eliminating the need for supervised data, annotations, and costly domain expertise.
- Multilingual experiments showcased ChatAgri's domain transferability, enabling quick adaptation to diverse agricultural applications. This crucial step accelerates the future development of General Purpose AI (GPAI).
- ChatAgri offers a cost-effective alternative to complex PLM-based methods for smart agricultural applications, with minimal hardware requirements.

² You can access ChatGPT by visiting the following URL: <https://chat.openai.com/chat> [Accessed on 2023.05].

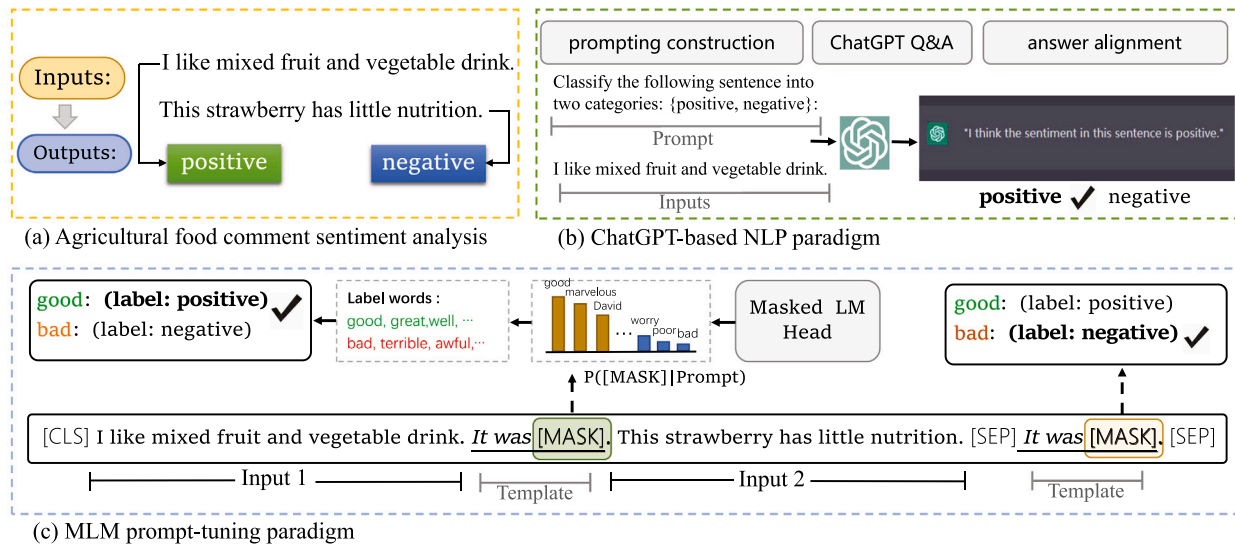


Fig. 2. The paradigm comparison of the ChatGPT-based NLP solutions and existing prompt learning paradigm using an agricultural sentiment analysis example. Part. (a) denotes the task prototype of the agricultural sentiment analysis; Part. (b) denotes the standard workflow of ChatGPT-based approaches; and Part. (c) denotes the standard workflow of Masked LM prompt-tuning methods.

- We have released the ChatAgri codes on GitHub³ to encourage further research in smart agricultural applications using ChatGPT.

The remainder of this paper is organized as follows: Section 2 provides an overview of the recent literature in related fields, with a focus on recent research for the agricultural text classification task, ChatGPT, and pre-trained language model-based NLP techniques. Section 3 presents a detailed description of the whole ChatAgri framework, including a detailed algorithmic description. In Sections 4 and 5, we conduct a comprehensive analysis of the comparison experiments between ChatAgri and several mainstream PLM-based methods, along with various ablated studies. Finally, Section 7 offers a concise summary of the primary contributions of our research and outlines future prospects for further sustainable smart agriculture development based on our findings.

2. Related work

In this section, we will review the related literature on accurately classifying cross-linguistic agricultural texts, recent advancements and applications in ChatGPT and its extensions, as well as PLM-based fine-tuning and prompt-tuning approaches in addressing the challenges of agricultural text classification.

2.1. Agricultural text classification

Over the past decade, the primary machine learning models (e.g. decision tree, CNN, LSTM, and GRU) [6] have been the dominant approaches in research on the agricultural document classification.

Azeez et al. [31] used the support vector machine (SVM) and decision tree induction classifiers to complete the regional agricultural land texture classification. Li et al. [32] simultaneously utilized the Bi-LSTM and the attention mechanism to further dynamically enrich the extracted multi-sources semantic features, which effectively improve the performance of agricultural text classification. Dunmon et al. [33] leveraged CNN to predict agricultural Twitter feeds from farming communities to forecast food security indicators, and demonstrated that

CNNs are widely superior to RNNs in agriculturally-relevant tweets sentiment classification.

Since the introduction of large models such as BERT [3] and GPT [34], many NLP tasks have achieved significant performance improvements and have gradually replaced traditional machine learning approaches [29]. Compared to traditional machine learning methods, large pre-trained language models are better equipped to handle the complexity scenarios, having received widespread attentions in both academic and industrial settings.

Shi et al. [35] employed BERT to identify the most representative information from unlabeled sources, which were manually labeled to construct the corpora of agricultural related news from diversified topics, enhancing the efficiency of labeling process and ultimately improving the corpora construction quality. Jiang et al. [1] automatically classify the French plants health bulletins to make these data easily searchable through fine-tuning BERT. Leong et al. [8,9] developed an automatic optical character recognition system for the categorization and classification of agricultural regulatory documents. To tackle the imbalance between the supply and demand of the agricultural market, Cao et al. [7] introduced a improved BERT-based sentiment analysis model for agricultural product evaluation through Internet reviews. The proposed BERT model with symmetrical structure accurately identifies the emotional tendencies of consumers, helping consumers evaluate the quality of agricultural products and helping agricultural enterprises optimize and upgrade their products.

2.2. Traditional machine learning methods, and PLM-based fine-tuning, and prompt-tuning

In the past, traditional machine learning approaches were commonly used for agricultural text processing. Xu et al. [36] propose agricultural exports time series-long short-term memory (AETS-LSTM), a novel method for predicting agricultural export fluctuations, achieving improved prediction performance. This helps agribusiness operators evaluate and adjust policies more effectively. Costa et al. [37] develop a knowledge-based system for identifying rice farming pests and disease symptoms, utilizing jaccard similarity coefficient (JSC) for critical information extraction and problem diagnosis.

Feature engineering-based methods had limitations in capturing the complexity and nuances of natural language, particularly in semantic complex situations [29]. PLMs [3,4,34,38] emerged as a powerful technique that revolutionized NLP, replacing many traditional methods [1,

³ Code has been released on Github: https://github.com/albert-jin/agricultural_textual_classification_ChatsGPT [Accessed on 2023.05].

33,39,40]. PLM-based fine-tuning became the mainstream learning technique in agricultural information processing [41], introducing additional network parameters and task-specific objective functions to fine-tune PLMs for downstream tasks. Cao et al. [7] improved BERT for sentiment analysis in agricultural product reviews, aiding consumer purchasing decisions. They fine-tuned BERT and its downstream classifier using TensorFlow. Jin et al. [19] proposed DictABSA, a dictionary knowledge-infused network, for sentiment analysis and agricultural text classification.

Nevertheless, PLM-based fine-tuned models may struggle with generalization and require substantial annotated data, hindering rapid development and easy deployment. Consequently, traditional PLM-based fine-tuning has diminished in NLP, giving way to a more promising paradigm called “prompt learning” or “prompt-tuning” [20]. Unlike fine-tuning, prompt-tuning utilizes constructed prompting templates to adapt downstream tasks to PLMs, excelling in few-shot or zero-shot scenarios [15,27,28,42]. It improves the performance of PLMs through optimized prompts and incorporates external knowledge from knowledge bases (KBs) to stimulate internal knowledge within PLMs [42].

2.3. ChatGPT

Before the emergence of ChatGPT technology, deep learning-based artificial intelligence methods played a significant role in various industries, including improving productivity, optimizing supply chain management, enhancing medical diagnostics, strengthening risk management, and achieving autonomous driving technology, among others. These representative works [43–45] in the field of deep learning have had a certain influence on the advent of ChatGPT.

Fang et al. [43] proposed an adaptive switching randomly perturbed particle swarm optimization algorithm (ASRPPSO) to optimize the initial cluster centroids of the FCM algorithm. They demonstrate the superiority of ASRPPSO over five existing PSO algorithms on benchmark functions. Zhang et al. [44] developed a novel dual-pathway-fusion-based sequence-to-sequence learning model (DPF-S2S) for text recognition in various environments. The model focuses on enhancing spatial information and extracting high-dimensional representation features to improve decoding accuracy. Extensive experiments validate the effectiveness of the proposed strategies in the DPF-S2S model. Wu et al. [45] presented a magnetic resonance imaging (MRI) oriented attention-based glioma grading network (AGGN). The network utilizes a dual-domain attention mechanism to incorporate both channel and spatial information, enabling the highlighting of key modalities and locations in the feature maps.

With the advancement of artificial intelligence and deep learning, ChatGPT, an advanced conversational technology in the general domain, has been introduced [14,17]. Developed by OpenAI, ChatGPT serves as an expert in all fields, possessing omnipotent and omniscient knowledge. ChatGPT, a disruptive revolution spanning various research domains, extends beyond NLP, offering a user-friendly interface for unprecedented access to the power of large language models. Also known as GPT-3.5, it acts as a conversational robot capable of understanding complex instructions and generating high-quality responses across diverse scenarios. With its significant contributions to multiple application scenarios, ChatGPT has revolutionized the role of virtual assistants. Compared to previous NLP models like BERT, BART, and T5, ChatGPT stands out with its extremely large size, prompting capabilities, and reinforcement learning from human feedback (RLHF).

As millions of users leverage language models, numerous use cases arise, unlocking ChatGPT's potentials. In empirical studies [29,30], ChatGPT excels in multilingual translations, particularly in high-resource languages. It performs comparably to Tencent TranSmart, DeepL Translate, and Google Translate. Impressively, ChatGPT aids in code debugging and generation [14]. Haque et al. [23] evaluate ChatGPT's ability to provide code snippets in Python, Java, and JavaScript, adhering to syntax and semantics. Bang et al. [46] use ChatGPT for

multimodal tasks like generating images from text, showcasing its superior code generation quality over other techniques.

The fascination with ChatGPT has led to investigations into its possibilities, especially in agriculture. Gao et al. [24] examine ChatGPT's feasibility for event extraction, highlighting its complexity and the need for comprehensive instructions. Wei et al. [26] develop ChatIE, a universal zero-shot information extraction framework using ChatGPT, handling tasks like named entity recognition, event extraction, and relation extraction. ChatIE employs a decomposed multi-stage approach, involving QA turns to discover element types and fill corresponding slots.

Moreover, OpenAI [14] launched GPT-4, an advanced large-scale multi-modal generative PLM in March, surpassing ChatGPT (GPT-3.5) in multi-modal image and text interaction, expanded character constraints, and enhanced semantic comprehension. GPT-4 holds immense promise for diverse applications and represents a major step towards achieving general-purpose technology. The official investigation of GPT-4 [13] confirms its substantial impact on occupations, particularly higher-wage occupations exposed to PLMs. A recent open letter, signed by renowned researchers, calls for a pause in “Pause Giant AI Experiments” to the development of GPT-5 due to perceived risks of GPT-4's formidable power [14]. The open letter has gained attention and even garnered support from OpenAI CEO Sam Altman, who recognizes the revolutionary and impressive future impact of General Purpose AI, exemplified by ChatGPT, across industries. As the capabilities of models like ChatGPT and GPT-4 continue to expand, it becomes crucial to carefully evaluate their deployment and ensure responsible and ethical use to mitigate any potential negative consequences.

3. ChatAgri: ChatGPT-based agricultural text classification

3.1. Methodology overview

Focusing on investigate the feasibility of applying ChatGPT to agricultural text classification, ChatAgri, one of the preliminary studies of ChatGPT-based agricultural applications is constructed in this paper, along with a series of systematically and exploratory experimental analysis discussed.

Through our investigations, there are no existing research works that systematically utilized ChatGPT to the text classification task until our ChatAgri proposed. To fill this gap, the question how to defined the corresponding general workflow for the ChatGPT-based agricultural text classification will be further discussed. Specifically, after referred to abundant latest literature, as shown in Fig. 3, we deem that almost all the ChatGPT-assisted applications can be divided into three phrases:

- Prompting Question Construction: The first stage which focuses on providing appropriate prompting strategies to be fed into ChatGPT;
- ChatGPT Q&A Inference: The second stage about the reasoning procedure of ChatGPT Q&A, which is transparent to us and can be regarded as a black box;
- Answer Normalization or Alignment: The third stage transferring the natural language intermediate response to the target label in the pre-defined categories.

Among these steps, in addition to the Q&A inference conducted by ChatGPT, a static reasoning procedure we cannot participate in modification, the prompting construction engineering and answer alignment engineering can be further optimized during our experiments. From a macro perspective, ChatAgri is a pipeline structure in which each procedure influence the final prediction performance to a certain extent, including the quality of constructed prompts, the selected ChatGPT version, and the priority of adopted answer mapping strategies. Thus, the next subsections will introduce multiple novel solutions which are utilized in our experiments to fully exert the enormous potential and superiority of the ChatGPT in ChatAgri.

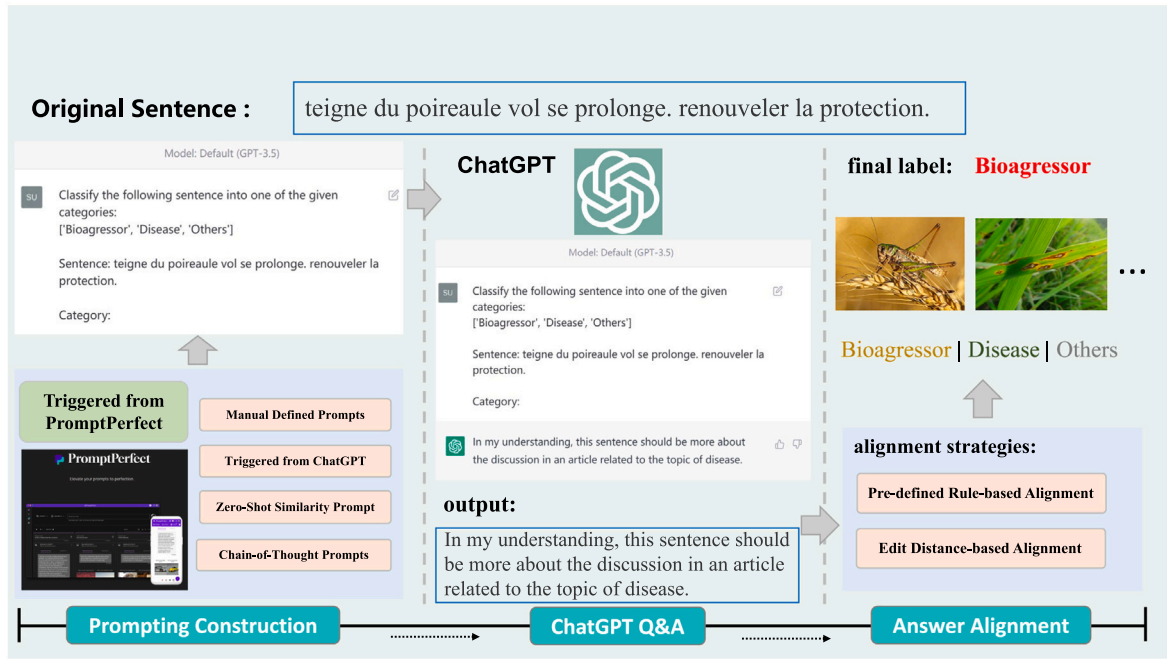


Fig. 3. The framework of ChatAgri, which is illustrated by an typical example in the agricultural natural disaster dataset, French Plant Health Bulletin. First (left), several prompting construction strategies were applied to generate prompts, and the ChatGPT question is constituted by integrating these prompts with the original sentence; Second (center), ChatGPT provides response based on the inputs; Finally (right), the answer alignment strategies were devised to classify the intermediate answer to pre-defined categories.

Furthermore, as opposed to the text classification in the universal domain, the agricultural text classification acted as a domain-specific research branch due to the additional requirements of domain expertise knowledge. Another crucial factor, domain-specificity, should also taken into more considerations and corresponding customized strategies.

The following chapters would successively elaborate the specific solutions during the entire experiments of ChatAgri.

3.2. Prompt question construction

It is widely acknowledge to us that prompting engineering is a cumbersome art that requires extensive experience and manual trial-and-errors [20,29]. To design the suitable prompts to trigger the sentence classification ability of ChatGPT, we investigate sufficient pioneering works that discuss about how to generate optimized ChatGPT prompting questions [24,25,47]. Specifically, as depicted in the left of Fig. 3, the adopted prompt generation strategies in this experiments includes: (1). manually defined prompts; (2). prompts triggered from ChatGPT; (3). prompts based on the zero-shot similarity comparisons; and (4). prompts based on Chain-of-Thought (CoT); These novel prompt generation strategies are discussed in the followings.

3.2.1. Manually defined prompts

Following the general communication habits, we manually elaborate several prompting templates, Table 1 displays the part of designed prompts. Note that it is necessary to provide ChatGPT with the two mentions: original textual context and pre-defined categories, through some appropriate ways. Furthermore, for simplicity, we insert two extra slots into the prompts to combine the corresponding mentions, which respectively are [SENT] (slot of sentence) and [CATE] (slot of categories).

To conduct the successive comparison experiments, we evaluate the specific effect of each candidate prompt to select the best candidate prompt. Formally, we employ a data sampling-based evaluation approach among these candidate prompts [46]. Concretely, we randomly selected a fixed number of samples (set as 100 during experiments by

Table 1

The partial manually devised prompts. [Res] denotes the response provided by ChatGPT.

No.	Prompting template
1	Classify the following sentence into one of the given categories: [CATE] \n Sentence: [SENT] \n Category: \t [Res]
2	Which categories do you think sentence: \n [SENT] \n belongs to, out of [CATE]? \n [Res]
3

Your task is to categorize the given sentence into one of the provided categories.

Please provide a clear and concise response that accurately identifies the category of the sentence to allow for categorizations.

The sentence is: [SENT] .

The categories are specified in the [CATE] .

The sentence to be classified is: {ChatGPT. Response} .

Fig. 4. The adopted prompt which is selected through the subset evaluation.

default) from the Twitter Natural Hazards dataset, then we further test the performance for each prompts on this subset by accuracy. After overall comparisons, the prompt which is shown in Fig. 4 is selected as the most suitable manually defined prompt for subsequent experiments.

Moreover, note that we add an extra command “Please only answer the category.” into prompts to ask ChatGPT not to generate redundant explanation around the ChatGPT reply, which might be a disrupting factors for subsequent text label decisions. The factor has also been taken into consideration for the subsequent prompting methods.

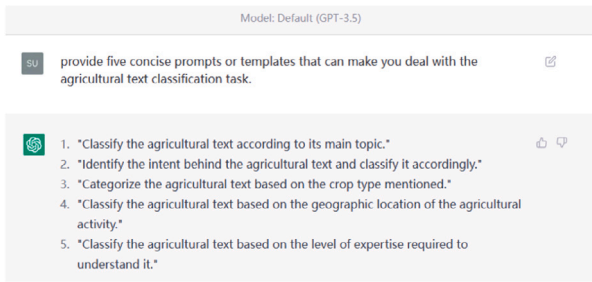


Fig. 5. Candidate prompt templates triggered by requests to ChatGPT (Model: GPT-3.5, Query Date: 2023.4.02).

3.2.2. ChatGPT triggered prompts

Drawing inspiration from the relevant literature [25,47], we posit that inquiring about ChatGPT itself could potentially yield valuable insights into the generation of high-quality templates. Thus, we seek inspiration from ChatGPT by asking ChatGPT with the recommendations for templates generation. Note that a similar preliminary study of Zhong et al. [25] suggests that the task-specific prompts can be triggered by using the following human inquiries:

> Provide five concise prompts or templates that can make you deal with the [x] task.

where the slot [x] means the specific task types. Experiments prove that this strategies performs well in most scenarios.

Correspondingly, as shown in Fig. 5, our request is intuitively constructed as follows:

> Provide five concise prompts or templates that can make you deal with the agricultural text classification task.

Afterwards, ChatGPT naturally answers us with several candidate responses, which is depicted in Fig. 5. The prompts that have been generated appear to be sensible and consistent in terms of their semantic content, while also exhibiting some noticeable distinctions in terms of their individual formats.

To this end, following the above described sampling-based evaluation method, we select the best-performed prompt to represent the ChatGPT triggered prompts for successive comparison experiments, which is shown as follows:

> Classify the agricultural text: [SENT] according to its main topic [CATE]."

3.2.3. Zero-shot similarity prompts

Motivated by previous few/zero-shot learning works that utilizes meta-learning paradigm [26,48], we devised a novel prompting strategies upon it, called zero-shot similarity-based prompting.

Typically, few-shot object classification is performed by leveraging sample and classifiers from similar classes by some distance measure and similarity functions, such as cosine similarity and squared ℓ_2 distance [48]. To give an example, let us consider the few-shot learning-based images classification task. Firstly, given an image to be classified, one extra representative image for each category was choosed. Then, they were embedded into the same low-dimensional space using an embedding network, such as siamese network, prototypical network, and matching network. Finally, the similarity threshold between the image to be classified and images from all-kind of categories is then used for label classification.

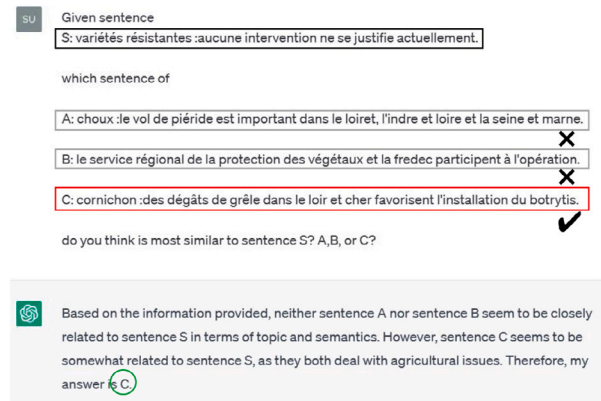


Fig. 6. The end-to-end direct similarity measurement QA-based prompting method for text classification.

Back to agricultural text classification, the adopted ChatGPT interface can be regarded as a special distance similarity measurement for evaluating the inter-relationship between two different sentences. All these procedures were conducted by performing one turn or multi turns QA. Specifically, we have designed two QA modes: end-to-end direct QA-based similarity evaluation and progressive comparison QAs-based similarity evaluation.

- **End-to-end direct QA-based:** Concretely, the most straightforward and simplest way is to directly ask ChatGPT that which sentence is most similar to the pre-classified sentence. Furthermore, we adopt the following prompt during experiments.

> Given sentence S: [SENT1], which sentence of A: [SENT2], B: [SENT3], . . . do you think is most similar to sentence S? A, B, . . . , or C?

In this manner, the text category can be finally determined. As see in Fig. 6, the target sentence can be classified to the category of sentence C based on only one-turn QA.

- **Progressive comparison QAs-based:** Similar to bubble sorting algorithm that compares pairs of elements at a time and subsequently applying the comparison to successive elements. Encouraged by the sorting algorithm, we incorporate its use in determining text similarity. Intuitively, we use the QA prompt:

> Given sentence S: [SENT0], which sentence A: [SENT1] and B: [SENT2] do you think is more similar to sentence S? Please answer using only A and B."

A typical example related to the three-classification problem was given in Fig. 7. Based on two-turn QAs, the target sentence can be classified to the category of sentence A based on the topic similarity comparison in the second QA stage. To our knowledge, we are the first to utilize the multi-stage similarity comparison approach to conduct the text classification task.

3.2.4. Chain-of-thought triggered prompts

In Jiao et al. s' [47] preliminary research of ChatGPT evaluation, they devised a *Pivot Prompting* translation strategy for ChatGPT-based multi-linguistic translator, which significantly improves the translation performance. *Pivot Prompting* translates source language to target language by using a high-resource pivot language (i.e. English by default) as a transition when two distant language is scarce. The above research

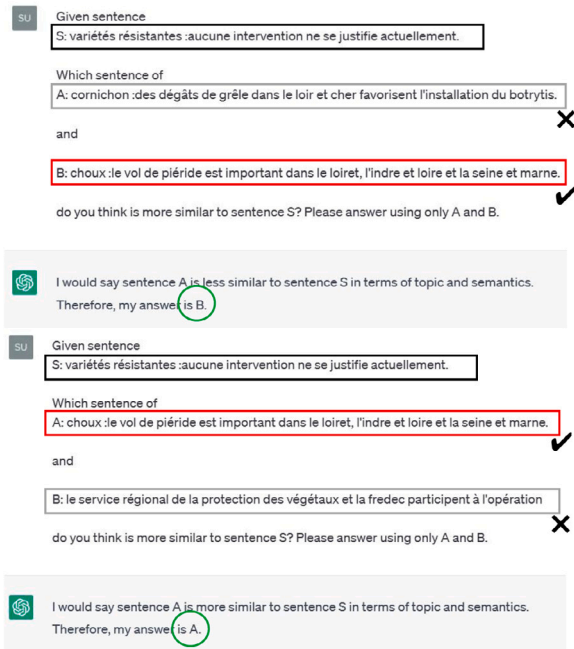


Fig. 7. The progressive similarity measurement QAs-based prompting method for text classification.

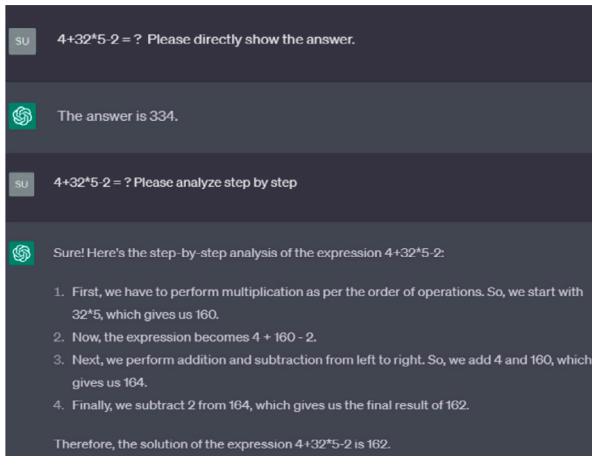


Fig. 8. The ChatGPT performance comparison between providing the answer directly and presenting a step-by-step calculation process in solving arithmetic problems. (Model: GPT-3.5, Query Date: 2023.3.15).

reflected that this intermediate transitional strategy is particularly effective in some special application scenarios. Jin et al.'s knowledge graph-based QA research [18] provides further evidence that these chains of reasoning are a critical factor that impacts the accuracy of the model.

Moreover, our inspection of ChatGPT's computational ability reveals that while ChatGPT tends to fall behind in its ability to reason and provide correct answers, it performs competitively when a step-by-step calculation process is used. Fig. 8 gives a typical example. To be more specific, while ChatGPT incorrectly provides the answer of 334 for the arithmetic problem $4+32 \cdot 5-2$, it is capable of correctly reasoning and arriving at the right answer for the same problem based on a step-by-step calculation process.

Building upon the experimental findings that support the effectiveness of step-by-step incremental reasoning, we explore the utility and viability of utilizing this technique for agricultural text classification.

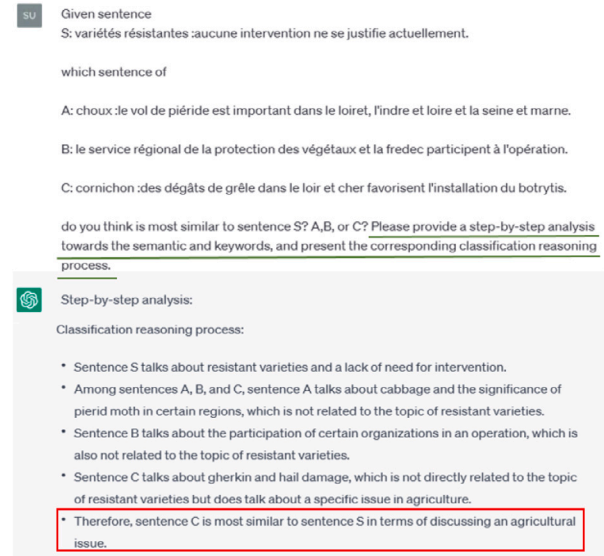


Fig. 9. The Chain-of-Thought-based prompting strategy which is built upon a simple and direct QA prompt. (Model: GPT-3.5, Query Date: 2023.3.15).

Concretely, we choose the *manually defined prompts* and *ChatGPT triggered prompts* as baselines. Also, drawing on these initial prompts, we require ChatGPT not only with delivering the final classification category but also with producing a corresponding comprehensive Chain-of-Thought reasoning analysis. For ease of illustration, as shown in Fig. 9, we further add the following expression based on the original QA prompt.

"> Please provide a step-by-step analysis towards the semantic and keywords, and present the corresponding classification reasoning process."

This section presents several feasible strategies which shares a diverse range of distinctive features that set them apart from one another. But the prompting engineering is more complex and nuanced than what we can observe at a superficial level, as it is influenced by multiple factors, with dataset features playing a particularly significant role. For example, experimental results indicated that the Chain-of-Thought triggered prompts performs particularly well on datasets with a high number of classification categories, but its effectiveness is not satisfactory when dealing with datasets with relatively simple classification (few categories), such as only two to three categories.

The upcoming experiments will systematically compare multiple prompting strategies proposed above to enable a comprehensive evaluation and research.

3.3. ChatGPT Q&A inference

ChatGPT is a state-of-the-art conversation robot which are based on the generative language model, Generative Pre-trained Transformer (GPT). During the question-answering phase, ChatGPT generates sequences using a probabilistic language model (PLM). The underlying architecture of ChatGPT, GPT-3.5, is a mainstream language probability model that has been widely used for natural language processing tasks.

The GPTs family uses the transformer architecture, which is a deep neural network that processes input data in parallel using multi-headed attention mechanisms. During the inference stage, the GPT model generates text by conditioning on a given prompt and sampling from a probability distribution of words that follow. The training process of ChatGPT involves training a neural network on a large corpus of

text data to learn the statistical patterns and relationships between words and phrases. The transformer-based architecture utilized in ChatGPT empowers it to effectively capture long-range dependencies in language. The output of the model at each time step depends on the previous tokens generated, creating a generative process that allows the model to generate coherent text.

In terms of commonalities with other PLMs, ChatGPT shares similarities with its predecessor models like GPT-3. The training process involves a large-scale supervised fine-tuning approach where the model is trained to predict the next word or phrase given the previous context. This method is known as “autoregressive” generation since each word is generated conditioned on the previously generated words. Furthermore, there are some differences in training generation between ChatGPT and other PLMs. One notable difference is the size of the model and the amount of training data. ChatGPT, particularly GPT-3, is one of the largest language models developed by OpenAI, with 175 billion parameters. This large model size allows for more complex representations and a better understanding of language. Another difference is the use of reinforcement learning in the training process. GPT-3 was trained using a combination of supervised fine-tuning and reinforcement learning. This reinforcement learning approach involved using a reward model to fine-tune the model’s responses based on user feedback, allowing it to improve over iterations.

In summary, the sequence probability generation formula used during the question-answering phase in ChatGPT is based on a mainstream language probability model. While it shares commonalities with other PLMs in terms of training generation, such as autoregressive generation and fine-tuning, it stands out due to its large model size, diverse training data, and the use of reinforcement learning in the training process.

Mathematically, the token generative procedure of ChatGPT can be represented as:

$$p(y|x) = \prod_{t=1}^T p(y_t|y_1, \dots, y_{t-1}, x) \quad (1)$$

where the \prod means the probability multiplication operator. Given the previous tokens y_1, \dots, y_{t-1} and the input prompt x , $p(y_t|y_1, \dots, y_{t-1}, x)$ is the probability distribution over the token y_t in t -th time step and T is the length of the generated sequence.

At this stage, we direct our focus towards ChatGPT and hypothesize that ChatGPT possesses inherent capabilities that enable it to act as an integrated zero-shot text classification interface through an interactive mode.

During the ChatGPT interaction process, we created a fresh conversation thread for each prompt to ensure that the previous conversation history would not impact ChatGPT’s responses. By adopting this methodology, ChatGPT is able to consistently exercise independent thinking and deliver optimal responses by leveraging the information provided by the user.

Besides applying the vanilla ChatGPT (GPT-3.5), our experiments also evaluated the capabilities of GPT-4 [14]. GPT-4 represents a new breakthrough in OpenAI’s ongoing efforts to advance the field of deep learning. The results showed that GPT-4 performed better than ChatGPT, even in some complex semantic text classification scenarios, as seen in the following section of related evaluations.

3.4. Answer alignment

After the above steps, using an appropriate prompt and ChatGPT for question-answering, ChatGPT provided feedback on the classification results for the corresponding text. Nevertheless, its unique characteristic of generating responses in a conversational way presents challenges for the subsequent analysis and evaluation of its outputs. Unlike traditional PLM-based text classification models, ChatGPT’s responses do not directly correspond to predefined labels, which means that an additional alignment strategy is required to convert these intermediate

answers into the final labels that can be used to calculate various performance metrics (e.g. accuracy and F1-score). We refer to this additional mapping strategy as the “answer alignment engineering”.

In our experiments, we investigated the impact of answer alignment engineering on the ChatGPT-based text classification’s performance. Specifically, we designed and implemented two different alignment strategies: rule-based matching strategy and similarity-based matching strategy. Both approaches involve a mapping process that maps the intermediate responses to the corresponding labels. The rule-based matching approach uses predefined rules to match the responses to the labels, while the string matching approach computes the similarity between the response and each label and selects the label with the highest similarity score.

- **Rule-based matching strategy:** Essentially, the rule-based matching strategy is a text matching method that involves using patterns or rules based on token attributes, such as part-of-speech tags, to match sequences of tokens in unstructured text data. During our experiments, we use the `Matcher`⁴ object in `spacy v3` to find the matched tokens in context to classify the sentence returned by ChatGPT. `spacy v3` is a leading industrial-strength natural language processing and analysis tool⁵ using Python. Specifically, we firstly analyze the text extraction patterns based on expert experience and ChatGPT’s historical output habits, and design and define a set of rules. Then, the rules are applied to the text data and the extracted information is verified and validated. Finally, after adjustment and optimization, a comprehensive set of matching rules is summarized;
- **Similarity-based matching strategy:** Although the former approach utilizes rigid matching with high accuracy, it is difficult to handle semantically ambiguous situations. To address this issue, we adopt the second strategy, which is the similarity-based matching strategy. Firstly, we aggregate and synthesize ChatGPT’s commonly expressed utterances under each category to establish a repository of pivot answers for each category. Subsequently, we apply the Levenshtein distance algorithm to compute the minimum edit distance between each pivot answer and the input answer being classified. The pivot answer with the smallest edit distance is regarded as the definitive category label. This approach offers comprehensive coverage and effectively mitigates the shortcomings of rule-based matching in accommodating ambiguous and nuanced language use. The string similarity-based matching strategy is depicted in Fig. 10.

In theory, neither of these two strategies can perfectly solve the problem of answer mapping. To overcome the challenge of answer mapping, we combined rule-based and similarity-based matching strategies in a pipeline approach. Specifically, we found that ChatGPT typically provides explicit category labels in natural language form. Therefore, in the first step, we tend to use the rule-based strategy to parse the intermediate answers. If the category is still uncertain, we then use the string similarity-based strategy to compute the similarity between the intermediate answer and each category’s answer examples, selecting the category with the highest similarity as the final classification. In our experiments, this approach can simultaneously improve the accuracy and recall rate of the answer mapping process effectively.

Nevertheless, this work mainly explored a character-based literal matching method that lacks semantic understanding. The method has certain limitations, whereas the deep neural network-based methods using PLMs are more adept at such scenarios. In our future work, we will attempt to use a PLMs-based semantic understanding model for this step, which theoretically can bring about better performance.

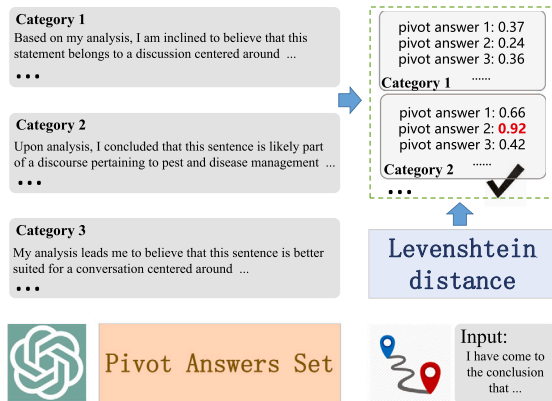
⁴ The `Matcher` tool is in <https://spacy.io/api/matcher> [Accessed on 2023.03].

⁵ `Spacy` can be accessed on <https://spacy.io> [Accessed on 2023.03].

Table 2

The statistical meta information of the adopted agricultural text classification datasets.

Dataset	Train samples	Test samples	Language	Categories	Label count
Amazon-Food-Comments	165863	16175	English	'negative', 'positive', 'neutral'	3
PestObserver-France	322	80	French	'Bioaggressor', 'Disease', 'Others'	3
Natural-Hazards-Twitter	45669	5074	English	'negative', 'positive'	2
Natural-Hazards-Type	5000	1000	English	'Hurricane', 'Wildfires', 'Blizzard', 'Floods', 'Tornado'	5
Agri-News-Chinese	52000	6500	Chinese	'Agricultural economy', 'Horticulture', 'Agricultural engineering', 'Farming', 'Fisheries', 'Forestry', 'Crops'	7

**Fig. 10.** The illustrating diagram of the similarity-based matching strategy.

4. Experimental setup

We perform a series of experiments in order to figure out exactly what kinds of factors of these devised strategies that indeed influence the final agricultural text classification performance of the ChatAgri in Section 5. Correspondingly, acting as a preliminary, this section mainly introduces the details of the experimental setups, including the used multi-linguistic datasets, the employed text classification baselines for model comparisons, the adopted evaluation metrics, and the adopted hyperparameters of our ChatAgri.

4.1. Datasets

To demonstrate the actual potentials of ChatAgri for classifying agricultural text, we carefully collect several suitable datasets for evaluation and validation, ranging from different types of categories (e.g. plant diseases, insect pests, and twitter natural hazards) and numbers of categories to different languages, including French, English, and Chinese. These datasets are respectively called Amazon-Food-Comments, PestObserver-France, Natural-Hazards-Twitter, Natural-Hazards-Type, and Agri-News-Chinese in our experiments, whose details are illustrated as follows.

- **Amazon-Food-Comments:** An amazon food comment dataset that contains nearly 200,000 positive samples, neutral samples, and negative samples, which can be used to perform text classification tasks for both positive, neutral, and negative reviews⁶;
- **PestObserver-France:** [1] A plant health bulletin classification dataset in French to estimate a agricultural prediction model that how well can it deal with heterogeneous documents and predict for natural hazards⁷;

- **Natural-Hazards-Twitter:** [49] A natural disaster dataset with sentiment labels of United States which is proposed to identify attitudes towards disaster response. It contains different natural disaster types and nearly 5,000 Twitter sentences⁸;
- **Natural-Hazards-Type:** In addition to recognize the sentiment polarities of Natural-Hazards-Twitter, we also re-organize it into a new disaster type classification dataset, denoted as Natural-Hazards-Type, to identify the natural disaster categories of text. Due to the large volume of the original Natural-Hazards-Twitter dataset, the new Natural-Hazards-Type dataset has taken a small subset of it, containing thousands of samples;
- **Agri-News-Chinese:** Besides the above existing datasets, we proposed a Chinese Agricultural short text classification dataset, namely Agri-News-Chinese, containing seven categories, such as agricultural economy and aquatic fishery. Its data source was collected and cleaned from the agricultural technology expert online system (ATE expert online system),⁹ with a total volume of approximately 60000 pieces of data, divided into the train and test sets by 9:1.

Table 2 gives a meta statistic for the five datasets, including the split distribution of train/test samples, the language scope, and the categories of textual topics.

4.2. Baselines

Existing extensive models for text classification can be divided into five major training paradigms: (1) traditional feature engineering-based machine learning (e.g. SVM, Decision Tree, and Random Forest) [31–33]; (2) word embedding-based deep learning (e.g. TextCNN, and TextRNN); (3) PLM-based fine-tuning, in which the PLMs include BERT [3], BART [4], T5 [5] and so on; (4) PLM-based prompt learning; and (5) the newest ChatGPT QA-based zero-shot learning paradigm that brought by ChatGPT recently (e.g. ChatIE [26], ChatEventExtract [24], and our ChatAgri).

To ensure the research comprehensiveness, the above introduced mainstream natural language understanding (NLU) paradigms were considered to be estimated and reported as the comparison baselines in our experiments. Specifically, besides the herein proposed ChatAgri, we adopted the following methods listed below for each mentioned learning paradigm.

- **SVM:** [31] Support Vector Machine (SVM) is a classic classification method pursuing maximization of support vector distance between multiple class hyper-planes for classification, typically in the text category classification task. SVM mainly classifies the text by calculating the unstructured discrete textual features, optimizing them into high-dimensional spatialized vector representations;

⁶ Access to <https://nijianmo.github.io/amazon/index.html> for more details of Amazon-Food-Comments [Accessed on 2023.02].

⁷ PestObserver-France can be downloaded from <https://github.com/sufianj/fast-camembert> [Accessed on 2023.02].

⁸ Natural-Hazards-Twitter can be downloaded from <https://github.com/Dong-UTIL/Natural-Hazards-Twitter-Dataset> [Accessed on 2023.02].

⁹ More details about ATE expert online system is available at <http://zjzx.cnki.net/> [Accessed on 2023.02].

- **Random Forest:** [31] Random Forest (RF) is also a well-known classification algorithm, belonging to the ensemble methods family, combines multiple weaker classifier to create a stronger classifier for categorical data;
- **TextCNN:** [50] Built on the top of pre-trained word vectors, TextCNN uses convolutional neural networks (CNN) as feature detector and utilizes kernels of different sizes to extract the valuable semantic feature for sentence classification. Lastly, the external softmax layer performs multi-classification on the convolutional logical values;
- **TextRNN:** [51] Based on pre-trained word embeddings, TextRNN integrates recurrent neural network (RNN) into the multi-learning framework. Specifically, TextRNN utilizes long short-term memory (LSTM) to address the issues of gradient vanishing and exploding, thereby resolving the challenge of capturing long-range dependencies within sequences;
- **BERT-based fine-tuning:** [3,7,29] Fine-tuning BERT has emerged as a widely employed methodology across diverse text processing tasks, including text classification. By generating contextualized word embeddings, BERT effectively captures both semantic and syntactic information associated with individual words. Leveraging its inherent strengths, BERT can be fine-tuned on specific tasks utilizing limited labeled datasets, rendering it a flexible and formidable solution for addressing an array of text processing objectives;
- **T5-based prompt-tuning:** [5,20,52] Different from the “pre-train then fine-tune” procedure of fine-tuning methods, the prompt-tuning paradigm induces those PLM to generate suitable target responses with the help of additional triggered sentences, which are called “prompts”. In prompt-tuning, the major research attention has been transferred on how to provide better prompts to activate the PLM’s rich internal prior knowledge. We use the PLM, Transfer Text-to-Text Transform (T5) to be the backbone. T5 is a unified very large PLM based on Transformer architecture, which converts all text processing tasks into Text-to-Text tasks;
- **BART-based prompt-tuning:** [4,52] We also investigate the usage of Bidirectional and Auto-Regressive Transformers (BART), being acted as the backbone for prompt learning. BART simultaneously incorporates the advantages of BERT and GPT (i.e. the characteristics of the context bidirectional modeling and the sequence joint probability hypothesis);

4.3. Evaluation metrics

In such agricultural text classification task that involves multiple label classification, accuracy and F1-score are two commonly used metrics.

Correspondingly, accuracy measures the proportion of correctly predicted samples among all predicted samples, is a simple and coarse-grained evaluation metric which only accumulates all the correct instances. And accuracy is calculated as follows:

$$Accuracy = Count_T / Count_N \quad (2)$$

where $Count_T$ represents the correctly predicted samples and $Count_N$ represents the total number of samples evaluated.

Comparatively, F1-score is considered to be a relatively fine-grained evaluation indicator than accuracy. In comparison to accuracy, F1-score is considered to be a higher confidence indicators which simultaneously considers the precision and recall. And F1-score is calculated as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad \text{where} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad \& \quad Recall = \frac{TP}{TP + FN}$$

In the equation presented, *Precision* and *Recall* refer to the precision and recall rate of the classification results, respectively. TP (true

positives) represents the number of samples whose actual and predicted class are both positive; FP (false positives) represents the number of samples whose actual class is negative but are predicted as positive; and FN (false negatives) represents the number of samples whose actual class is positive but are predicted as negative.

Specifically, F1-score includes several calculating strategies: micro-F1, macro-F1, and weighted-F1. Without considering micro-F1 and macro-F1, we utilize the weighted-F1 as it accounts for the classification performance of categories under varying weights, thereby providing greater reference value.

4.4. Hyperparameter settings

During our experimental procedure, there are various meta settings for all kinds of hyperparameters. The optimal hyperparameters, determined by their superior performance on the development set, will be selected for the final evaluation. The meta settings are summarized as follows.

We adopted the pretrained word vectors, GloVe [53], as the embeddings of the baselines of TextCNN and TextRNN. GloVe leverages the word co-occurrence statistics that can capture both syntactic and semantic relationships between words.¹⁰ Considering the trade-offs between computational limitations and performances and to ensure experimental competitiveness and stability, we adopted the version “bert-base-uncased”¹¹ for the PLM BERT, the version “t5-base”¹² for the PLM T5, and the version “facebook/bart-base”¹³ for the PLM BART respectively. The code implementation is developed using Python 3.7¹⁴ and PyTorch 1.9.0¹⁵ frameworks. For experimental simplicity, the prompts of the prompt-tuning baselines are pre-defined as “Given a sentence of [SENT], it is more like to be a topic of {SLOT} from [CATE]”, and the probability scores of the estimated words in the position of {SLOT} are then regarded as the intermediate answers for the final classification. Furthermore, the experimental hardware environment comprises a CPU Intel Core i9-9900k, and a single Nvidia GPU of GTX 1080Ti.

5. Experimental results and analyses

Next, we conducted a series of baseline comparison experiments and ablation experiments to analyze and explore the specific connections between various key factors that affect the performance of ChatAgri on agricultural text classification tasks. We first verified the competitiveness and superiority of ChatAgri relative to known state-of-the-art (SOTA) models. Then, we systematically investigated the impact of different prompting strategies on the classification accuracy for text classification. Moreover, we also attempted to apply GPT-4 and investigated the superiority of GPT-4 compared to the basic version of ChatGPT, GPT3.5. The systematic analysis towards extensive empirical results firmly demonstrate the enormous potentials, feasibility, and broad application prospects of ChatGPT in agricultural text classification tasks.

¹⁰ GloVe embedding can be downloaded from: <https://nlp.stanford.edu/projects/glove/> [Accessed on 2023.03].

¹¹ BERT can be obtained from: https://huggingface.co/docs/transformers/model_doc/bert [Accessed on 2023.03].

¹² T5-base can be obtained from: <https://huggingface.co/t5-base> [Accessed on 2023.03].

¹³ BART can be obtained from: https://huggingface.co/docs/transformers/model_doc/bart [Accessed on 2023.03].

¹⁴ Python can be downloaded from: <https://www.python.org/downloads/release/python-370> [Accessed on 2023.03].

¹⁵ Pytorch can be downloaded from: <https://pytorch.org/blog/pytorch-1.9-released> [Accessed on 2023.03].

Table 3

Performance Statistics of all baselines and ChatAgri on all adopted datasets. We respectively boldface and underline the score with the best performance and the second-best performance across all models (**Query Date: 2023.3.16**).

Learning paradigms	Baseline methods	Amazon-Food-Comments		PestObserver-France		Natural-Hazards-Twitter		Natural-Hazards-Type		Agri-News-Chinese	
		acc	weighted-F1	acc	weighted-F1	acc	weighted-F1	acc	weighted-F1	acc	weighted-F1
Traditional Machine Learning	SVM	0.627	0.624	0.672	0.655	0.763	0.742	0.811	0.811	0.523	0.522
	Random Forest	0.647	0.643	0.664	0.652	0.787	0.755	0.863	0.863	0.553	0.534
Word Embedding-based learning	TextCNN	0.748	0.742	0.715	0.704	0.834	0.816	0.914	0.914	0.792	0.785
	TextRNN	0.727	0.725	0.707	0.697	0.845	0.827	0.931	0.931	0.812	0.801
PLM-based fine-tuning	BERT-based fine-tuning	0.767	0.764	0.736	0.714	0.869	0.839	0.945	0.945	0.826	0.819
	T5-based prompt-tuning	0.805	0.798	<u>0.764</u>	0.753	<u>0.874</u>	<u>0.857</u>	0.966	0.966	0.859	0.854
PLM-base prompt-tuning	BART-based prompt-tuning	<u>0.800</u>	<u>0.795</u>	0.757	<u>0.767</u>	0.875	0.865	<u>0.971</u>	<u>0.971</u>	0.867	0.862
ChatGPT-based Prompt QA	ChatAgri-base (Ours)	0.798	0.793	0.794	0.789	0.866	0.853	0.978	0.978	<u>0.863</u>	<u>0.856</u>

5.1. Methods comparison

Table 3 details comprehensive experimental results on the agricultural text classification task for our model ChatAgri and existing state-of-the-art approaches. In this table, as shown by multiple rows before the row data of *ChatGPT-based Prompt QA*, we conducted a systematic evaluation of the classification performance of these baseline models on these five datasets based on the above described hyperparameter settings. The time node of ChatGPT interface calls is March 16, 2023. Subsequent OpenAI official updates may lead to certain performance fluctuations towards the ChatGPT interface. The last row shows the evaluation results of our ChatAgri. For simplicity and clarity, we took the primary designed solution of ChatAgri as the basic model of ChatAgri for comparison. Specifically, we used the manually defined prompts, which is illustrated in Section 3.2.1, as the prompting template for ChatAgri. And we simultaneously adopted the rule-based and similarity-based text pattern matching strategy for the answer alignment engineering. Correspondingly, we labeled this basic model of ChatAgri as **ChatAgri-base**.

In Table 3, we classified all the existing agricultural text classification methods explored in this experiment according to their belonged learning paradigms. Among them, these methods based on fine-tuning PLM and PLM prompt engineering can be seen as the latest optimal benchmark approaches, and are respectively recorded in the last few rows of the table. From the table, it can be clearly observed that our ChatAgri has achieved exciting and competitive performance on some specific datasets, such as PestObserver-France and Natural-Hazards-Type. Not to mention surpassing traditional machine learning methods or word vector-based representation learning methods by an absolute gap of over 10% to 20%, which is a noticeable performance margin. Compared with the latest Transformer PLM-based deep learning methods, ChatAgri is also a particularly strong presence, with no loss in accuracy or weighted-f1 compared to these SOTA methods. Specifically, ChatAgri significantly outperformed the PLM-based fine-tuning method represented by fine-tuned BERT by about 3.0% accuracy on the PestObserver-France dataset, and outperformed the PLM-based prompt-tuning method represented by prompt-tuned BART by approximately 2.2% weighted-f1 indicator. Similarly, ChatAgri also surpassed the above two state-of-the-art models by 0.6% accuracy and weighted-f1 indicators on the Natural-Hazards-Type dataset. In addition, the performance of ChatAgri on other datasets is also impressive. For example, it can be seen from the table that the performance of ChatAgri on the Agri-News-Chinese dataset have significantly surpassed the PLM-based fine-tuning method represented by fine-tuned BERT by about 3.7% accuracy and 4.7% weighted-f1 indicator. In addition, ChatAgri's performance is also slightly higher than the PLM-based fine-tuning method represented by prompt-tuned T5 by approximately 0.4% accuracy and 0.2% weighted-f1.

In addition, we further explored the reasons why ChatAgri performed more strongly on some datasets but slightly worse than previous SOTA methods on others. By observations from Table 3, we found that ChatAgri had obvious advantages on two minority language datasets, PestObserver-France and Agri-News-Chinese, but performed poorly on the widely-used English datasets, Amazon-Food-Comment and Natural-Hazards-Twitter. We speculate that this is mainly due to the difference in the scale of large-scale language corpus training for different languages. After comprehensive investigations on latest literature [13,14,46], we can conclude that ChatGPT excels at handling various cross-linguistic tasks. Unlike previous methods based on traditional PLMs, ChatGPT's learning corpus is totally comprehensive and of high quality, covering the majority of languages spoken in most countries. Moreover, ChatGPT's ultra-large parameter size allows it to memorize and master more linguistic knowledge, not just limited to English. Therefore, in terms of cross-lingual understanding capability, ChatGPT is significantly superior to traditional PLM models (e.g. BERT, RoBERTa, and BART). Correspondingly, traditional PLM models perform poorly on less commonly spoken language datasets, as their learning corpus is far less comprehensive and of lower quality than that of ChatGPT. This probably is the primary factor that allows ChatAgri to perform well on various minority language datasets regardless of these datasets' linguistic characteristics.

On the Natural-Hazards-Type disaster category classification dataset based on the transformation of Natural-Hazards-Twitter, we found that both the PLM-based method and ChatAgri performed very well, fluctuating around 94% to 97% of accuracy and weighted-f1, which meets almost all the users' needs. By observing this dataset itself, we observe that most of the text in the dataset can be classified by using some fixed phrases as trigger words. For example, there is a sentence in the dataset: "*Florida governor declares state of emergency ahead of Dorian and warns Floridians on the East Coast*", where the word "*Dorian*" essentially belongs to the topic of a happened American hurricane disaster. As we know, a simple semantic context always can make the training and prediction of NLU tasks much simpler, so these existing SOTA models have achieved satisfactory performances. It is worth mentioning that during the process of reorganizing the Natural-Hazards-Twitter dataset into the Natural-Hazards-Type dataset, we intuitively maintained the same quantity of test samples for each category. Therefore, the calculation results of the accuracy indicator on the Natural-Hazards-Type dataset are the same with the weighted-F1 indicator.

The above discussion fully demonstrate the superiority of ChatGPT in agricultural text classification: even though ChatGPT has not been trained on any training set, it can still outperform all kinds of SOTA methods that trained on large-scale training sets. Note that ChatAgri-base used as a comparison baseline here solely employs the manually defined prompting strategy, which is a basic and simple one. Even the simple ChatAgri can achieve impressive results, which makes us

Table 4

Comparative experimental results of ChatAgri-base and various model variants of ChatAgri that utilized various advanced prompts, where the ChatAgri-base can be regarded as a basic ChatAgri implementation (Query Date: 2023.3.24).

Prompting strategies	Amazon-Food -Comments		PestObserver -France		Natural-Hazards -Twitter		Natural-Hazards -Type		Agri-News -Chinese	
	acc	weighted -F1	acc	weighted -F1	acc	weighted -F1	acc	weighted -F1	acc	weighted -F1
Manually Defined Prompts (ChatAgri-base)	0.798	0.793	0.794	0.789	0.866	0.853	0.978	0.978	0.863	0.856
ChatGPT Triggered - Prompts	0.806 ↑ 0.8%	0.803 ↑ 1.0%	0.815 ↑ 2.1%	0.812 ↑ 1.4%	0.871 ↑ 0.5%	0.862 ↑ 0.9%	0.978 = 0.0%	0.978 = 0.0%	0.874 ↑ 1.1%	0.867 ↑ 1.1%
Zero-Shot Similarity - Prompts	0.810 ↑ 1.2%	0.807 ↑ 1.4%	0.824 ↑ 3.0%	0.821 ↑ 2.2%	0.874 ↑ 0.8%	0.866 ↑ 1.3%	0.975 ↓ 0.3%	0.975 ↓ 0.3%	0.863 = 0.0%	0.856 = 0.0%
Chain-of-Thought Triggered - Prompts	0.816 ↑ 1.8%	0.814 ↑ 2.1%	0.832 ↑ 3.8%	0.829 ↑ 3.0%	0.874 ↑ 0.8%	0.866 ↑ 1.3%	0.981 ↑ 0.3%	0.981 ↑ 0.3%	0.889 ↑ 2.7%	0.883 ↑ 2.7%

more convinced that the ChatGPT-based solution will be the future direction for the continuous research development of agricultural text classification.

5.2. Improving ChatGPT with advanced prompting strategies

In order to explore the influence of different prompt generation strategies to the final classification performance, we conducted systematic evaluations and in-depth explorations of various prompt generation strategies introduced in Section 3.2 to clarify the advantages and significance of different prompt generation strategies in this section. The current date for ChatGPT interface calls is March 24, 2023. Subsequent OpenAI updates to the ChatGPT official API may influence the future function calls, leading to certain performance discrepancies.

From the first two rows of Table 4, it can be discovered that the ChatAgri which adopts ChatGPT Triggered-Prompts outperforms the Manually Defined Prompts strategy counterpart in most cases, indicating that ChatGPT can generate better prompts to trigger its more comprehensive knowledge for more accurate prediction. For instance, ChatAgri based on ChatGPT Triggered-Prompts improved the accuracy by average 2.1% and 1.1% on the PestObserver-France and Agri-News-Chinese datasets, respectively, compared to ChatAgri based on Manually Defined Prompts. This empirically demonstrates that prompt engineering for ChatGPT should be combined with ChatGPT's own understanding and feedback to achieve better classification performance.

From the third and fourth rows of Table 4, it can be observed that the Zero-Shot Similarity-Prompts strategy performs significantly better than the baseline prompts on the first three datasets, but its performance on the Natural-Hazards-Type and Agri-News-Chinese datasets is relatively unsatisfactory, even falling behind the basic prompts, namely Manually Defined Prompts and ChatGPT Triggered-Prompts. For example, ChatAgri based on Zero-Shot Similarity-Prompts reduced the accuracy and weighted-f1 by 0.3% compared to ChatAgri-base based on Manually Defined Prompts on the Natural-Hazards-Type dataset.

We can also easily observe from Table 4 that the Chain-of-Thought Prompts strategy significantly improves the overall task performance on all datasets, and its effect is better than that of ChatAgri based on Zero-Shot Similarity-Prompts. Especially on the Natural-Hazards-Type and Agri-News-Chinese datasets, Chain-of-Thought Triggered-Prompts has further improved, which is an excellent effect that Zero-Shot Similarity-Prompts cannot achieve. For example, on the Agri-News-Chinese dataset, Chain-of-Thought Triggered-Prompts simultaneously improved the accuracy and weighted-f1 by average 2.7% compared to ChatAgri-base.

It is worth mentioning that for the binary classification dataset Natural-Hazards-Twitter, the classification process based on the Chain-of-Thought rules only requires one comparison step, and the pivot sentence selected by this strategy is exactly the same as that used by Zero-Shot Similarity-Prompts. Therefore, the performance of the Chain-of-Thought Prompts and Zero-Shot Similarity-Prompts strategies is the same here. Moreover, due to the simple semantics of the Natural-Hazards-Type constructed by us, the prediction effect of

various ChatAgri model variants is close to saturation. Therefore, the Natural-Hazards-Type dataset is not more persuasive than other datasets in terms of reference value.

In summary, Chain-of-Thought Triggered-Prompts is particularly good at handling texts with many classification categories in multi-classification tasks, which also confirms the effectiveness of the divide-and-conquer idea of splitting complex multiple classification tasks into multiple simple binary classifications for handling slightly complex classification tasks. In contrast, Zero-Shot Similarity-Prompts performs relatively poorly when there are many classification categories, and even worse than the effects of Manually Defined Prompts and ChatGPT Triggered-Prompts. We speculate that the main reason is that the selection of pivot sentences is not perfect on the one hand, and on the other hand, when ChatGPT judges the specific similarity of multiple semantically similar pivot sentences, multiple semantically similar pivot sentences can easily confuse ChatGPT, leading to its easy misjudgment of the final classification result.

5.3. Few-shot prompt-tuning and zero-shot ChatAgri

Although most representative text classification methods are based on supervised learning with a large volume of high-quality annotated samples. The fact is, the annotation procedure of supervised corpora demands the expertise of domain specialists and is expensive and time-consuming, as well as a significant amount of manual efforts. Thus, in specific practical application scenarios, it is often more widespread and ubiquitous to apply data-scarce learning due to insufficient resource and scarce data.

As numerous literature have suggested [20,27,28], prompt-learning is particularly useful in data insufficient scenarios. It is a powerful and promising NLP technique which fully leverages the prior knowledge learned from the PLM's pre-trained stage. By using the prompting tricks, prompt-learning allows PLMs quickly adapt to various new tasks while learning on a small amount of data. Here, we delved in-depth into the characteristics, differences, and interactions between ChatGPT and prompt-learning paradigms. The evaluation statistic of these prompt learning methods was simulated based on the open-source framework *OpenPrompt*. *OpenPrompt* [52] is an advanced research toolkit developed by Tsinghua University.¹⁶ *OpenPrompt* integrates various prompt-based learning methods, making it easy and feasible for researchers to quickly develop and deploy their prompt-tuning solutions.

Correspondingly, we provided a detailed comparison to explore the relationships between ChatAgri and PLM-based prompt-tuning methods under few-shot and zero-shot learning settings. As shown in Table 5, we report the experimental results of these SOTA methods (i.e. T5-based prompt-tuning, BART-based prompt-tuning and ChatAgri) under the few-shot learning and zero-shot settings.

Here, we briefly introduce the reasons why the prompt-tuning method tends to use T5 and BART instead of BERT. As we know,

¹⁶ OpenPrompt can be accessed at <https://github.com/thunlp/OpenPrompt/> [Accessed on 2023.03].

Table 5

Performance statistics of ChatAgri and prompt learning baselines in the zero/few-shot supervised learning. Values (%) in **green** represent the increased performances of ChatAgri (zero-shot) compared to the second-best results (50-shot).

Few-shot learning	Methods	Amazon-Food-Comments		PestObserver-France		Natural-Hazards-Twitter		Natural-Hazards-Type		Agri-News-Chinese	
		acc	weighted-F1	acc	weighted-F1	acc	weighted-F1	acc	weighted-F1	acc	weighted-F1
Zero-Shot	T5-based prompt-tuning	0.521	0.523	0.474	0.466	0.562	0.545	0.597	0.597	0.425	0.419
	BART-based prompt-tuning	0.545	0.539	0.439	0.431	0.573	0.566	0.639	0.639	0.452	0.447
20-Shot	T5-based prompt-tuning	0.605	0.595	0.585	0.578	0.674	0.651	0.757	0.757	0.563	0.559
	BART-based prompt-tuning	0.627	0.609	0.563	0.554	0.643	0.626	0.761	0.761	0.594	0.592
50-Shot	T5-based prompt-tuning	0.679	0.674	<u>0.656</u>	<u>0.647</u>	0.732	0.719	0.831	0.831	<u>0.766</u>	<u>0.760</u>
	BART-based prompt-tuning	0.694	<u>0.688</u>	0.643	0.629	<u>0.758</u>	<u>0.746</u>	<u>0.854</u>	<u>0.854</u>	0.742	0.738
Zero-Shot (Default)	ChatAgri-base (Ours)	0.798 ↑10.5%	0.793 ↑10.5%	0.794 ↑15.1%	0.789 ↑16.0%	0.866 ↑10.8%	0.853 ↑10.7%	0.978 ↑12.4%	0.978 ↑12.4%	0.863 ↑12.1%	0.856 ↑11.8%

BERT learns contextualized representations by considering both left and right contexts, while T5 and BART are text-to-text transfer transformer models trained on a diverse range of tasks. BERT excels in fine-tuning with its bidirectional context understanding, while T5 and BART are flexible and adaptable due to their text-to-text format. Prompt-tuning typically involves fine-tuning the “autoregressive” Transformers by providing explicit prompts or instructions during the training process. The prompts guide the model to produce desired outputs for specific tasks. This approach is particularly useful when there is a need for fine-grained control over the model’s behavior or when the downstream task requires explicit instructions. Thus, prompt-tuning fine-tunes T5 and BART models with explicit prompts, useful for fine-grained control and explicit task guidance. BERT is commonly used for fine-tuning, while T5 and BART are leveraged in prompt-tuning for their adaptability. Both models have their own advantages and are applied based on the specific requirements of the downstream tasks.

Specifically, from the first row of Table 5, it can be seen that prompt learning methods are extremely effective in zero-shot learning (i.e., without any training on any samples), far surpassing the performance of models that guess based on average probability. For instance, on the Natural-Hazards-Twitter dataset, the BART-based prompt-tuning method achieved an accuracy of 57.3% in zero-shot learning, compared to a performance of 33.3% based on average probability, an improvement of about 24 percentage points. Especially on the five-classification dataset, Natural-Hazards-Type, the evaluated accuracy was 63.9%, which is much higher than the baseline accuracy of 20% for random prediction. In addition, under the 20-shot and 50-shot few-shot settings, the improvement of these prompt learning methods is even more significant, and the specific experimental results can be found in the third and fourth rows. The above statistical results indicate that prompt learning methods are very effective in training with small amounts of data.

Most impressively, it can be obviously observed from the table that ChatAgri performs significantly better than these prompt learning methods and achieves state-of-the-art performances in most aspects, regardless of different classification category topics and counts. The text classification performance of ChatAgri-base has surpassed these SOTA models in all test datasets with a significant improvement, demonstrating its superiority in all aspects. For example, compared with the baseline BART-based prompt-tuning that trained on 50-shot setting, ChatAgri-base yielded approximately absolute 10.5%, 15.1%, and 10.8% improvements in accuracy on datasets Amazon-Food-Comment, PestObserver-France, and Natural-Hazards-Twitter, respectively. It goes without saying that even compared to prompt learning models under zero-shot learning, those better performed, which is trained on a small amount of data, are significantly inferior to the ChatGPT-based classification framework ChatAgri without any fine-tuning. In addition,

better prompt engineering, ChatGPT models, and answer alignment engineering could further bring better results to the ChatAgri technology. Overall, ChatAgri has essentially surpassed the existing state-of-the-art prompt learning paradigm in all aspects, which is also the enormous potentials brought by the ultra-large-scale models.

In conclusion, ChatAgri shows its effectiveness and superiority in data-insufficient learning scenarios, indicating that ChatGPT has strong cross-domain and generalization capabilities. This kind of generalization is one of the directions for the development of future General Purpose AI, as it can help us build more flexible and adaptable intelligent systems that can handle various tasks and scenarios.

As we know, better performance would like to be obtained once using smoother prompts or update ChatGPT itself. As the impact of advanced prompting strategies has been investigated in Section 5.2, we then explore the potentials of upgrading the ChatAgri framework with more advanced ChatGPT, GPT-4.

5.4. Potentials between ChatGPT and GPT-4

Just as we were conducting research on vanilla ChatGPT (GPT-3.5) in March to April, 2023, OpenAI coincidentally released their latest powerful conversational system, GPT-4 [14], which serves as an improved version of ChatGPT. Thus, it is necessary to conduct additional exploration experiments to evaluate the overall performance of GPT-4, the upgraded ChatGPT, in the agriculture field text classification task.

Building on the advanced technologies learned from ChatGPT, GPT-4 has been iteratively refined to achieve unprecedented levels of authenticity, controllability, and rejection of undesirable outputs. In terms of model parameter scale, GPT-4 is expected to have over 1 trillion parameters, a significant increase from the GPT-3.5’s 175 billion parameters. This means that GPT-4 will be able to handle larger amounts of data and generate longer, more complex, coherent, accurate, diverse, and creative text. In terms of overall capability, compared to the previous version of ChatGPT, GPT-4 boasts improved performances in advanced reasoning, handling complex instructions, and demonstrating more creativity.

But GPT-4 currently has a cap of 25 messages every three hours by the latest released policy of OpenAI. It is the computation resource scarcity that caused the limited API capacity, which is far way from reaching the demand of the comprehensive experiments towards GPT-4 based ChatAgri. To overcome those pitfalls, we have taken a relatively balanced approach based on the trade-offs between experimental effectiveness and resource consumption (running time and empirical cost) in our experiments. Specifically, we made several reasonable reductions to the experiment from three perspectives: the linguistic categories, scales and their contributions of the datasets. The specific adjustments and arrangements for this experiment are as follows:

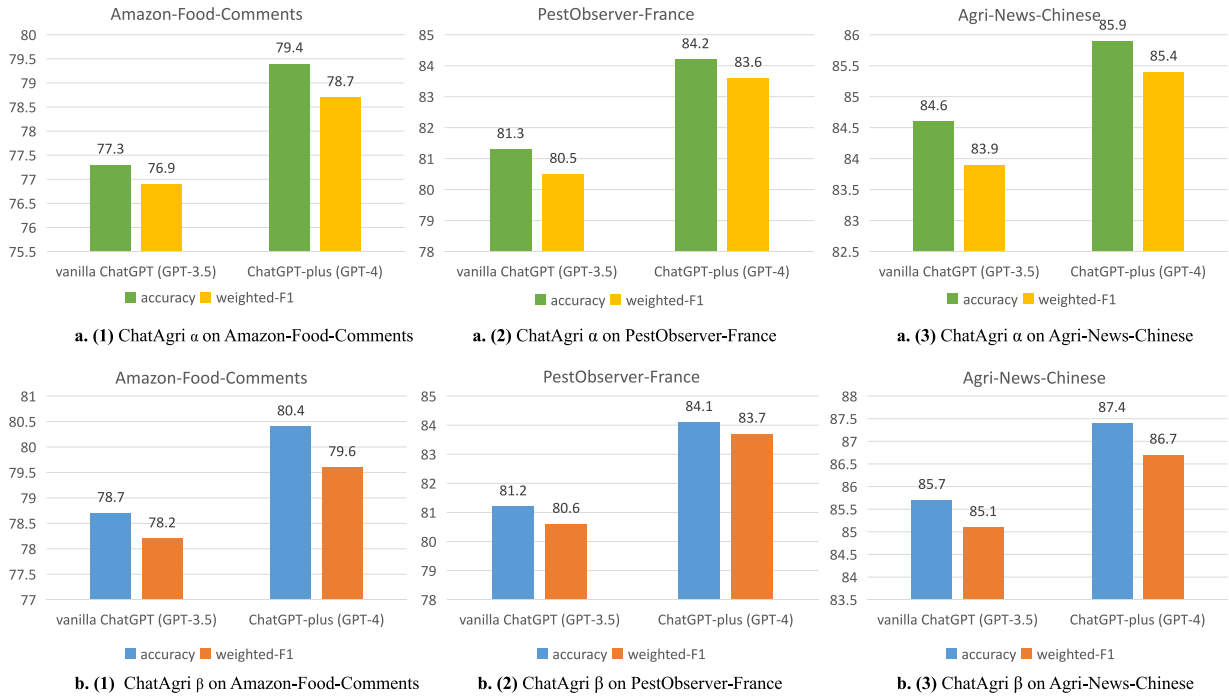


Fig. 11. The values shows the absolute metrics of accuracy and weighted-F1, which are reported using (%). The first group of a.(1), a.(2) and a.(3) denotes the ChatAgri $_{\alpha}$, and the second group of b.(1), b.(2) and b.(3) denotes the ChatAgri $_{\beta}$ counterpart. Reported results were averaged over 5 runs to ensure experimental reliability and robustness.

- For dataset selection, in order to comprehensively evaluate the performance of cross-linguistic text classification tasks, we selected three datasets that represent English, Chinese, and French contexts: Amazon-Food-Comments, PestObserver-France, and Agri-News-Chinese;
- For the specific samples to be evaluated, for each independent experiment, we randomly selected 100 samples from the original evaluation set as the evaluation subset;
- For the selection of the baselines, we used two ChatAgri models based on manually defined prompts and prompts triggered from ChatGPT, respectively labeled as ChatAgri $_{\alpha}$ and ChatAgri $_{\beta}$;
- To ensure the reliability and accuracy of the experimental results, we conducted 5 rounds of random screening and corresponding evaluations for each dataset, and took the average of the results from multiple rounds as the final evaluation result.

According to a series of comparative experiments, we found that GPT-4 performs better than vanilla ChatGPT, GPT-3.5. Specifically, as illustrated in Fig. 11, from which we can observe that the overall performance of ChatAgri $_{\alpha}$ and ChatAgri $_{\beta}$ equipped with GPT-4 is better than the counterparts equipped with vanilla ChatGPT. For example, as shown in a. (2) of Fig. 11, the GPT-4 based ChatAgri $_{\alpha}$ overwhelmingly outperforms the GPT-3.5 based ChatAgri $_{\alpha}$ by obtaining about 2.9% and 3.1% absolute gains of accuracy and weighted-F1 on the PestObserver-France dataset. As shown in the second group, GPT-4 also has brought a significant performance gain to ChatAgri $_{\beta}$ when compared with the vanilla ChatGPT-equipped counterpart on both the Amazon-Food-Comments and Agri-News-Chinese datasets, by achieving averaged 1.7% absolute accuracy gains. These experiment results powerfully demonstrate that GPT-4 can further exert its potentials and gain a better semantic understanding capability in handling the agricultural text classification task.

Especially in some complex semantic scenarios, like a semantic context containing a large number of semantically similar but subtly different texts, the classification accuracy of GPT-4 is significantly higher than that of vanilla ChatGPT. These results indicate that GPT-4 has higher accuracy and robustness in handling complex semantic

texts, and has a wider range of application prospects. Overall, the performance of GPT-4 is proved to be much superior and more stable than the vanilla ChatGPT. So far, we sincerely hope that in the future, OpenAI will provide greater support for the successive GPT series, including GPT-4 and even more advanced versions, so that we can fully leverage the benefits brought by advanced General Purpose AI in all aspects of future sustainable agricultural applications.

6. Case study

In this section, we provide a more in-depth exploration of ChatAgri's workflow details and investigate the key factors that influence its results. We thoroughly analyze a series of typical intermediate cases to understand the standard workflow of the designed answer alignment strategy and highlight its distinctive features. Additionally, by presenting several representative instances of ChatGPT's intermediate answers, we showcase the impressive cross-linguistic capabilities of ChatGPT in fine-grained semantic understanding. This allows us to better understand the response characteristics of ChatGPT in different scenarios.

6.1. Workflow features of answer alignment strategy

Here, we collected the statistical features of ChatGPT's responses to agricultural text classification, further exploring the response patterns of ChatGPT to the text prompts we designed. Based on this, we analyzed the adaptability of the answer alignment strategy to the response characteristics of ChatGPT. For experimental simplicity, the case analysis conducted here involves using *manually defined prompts*, as described in Section 3.2.1, to trigger responses from ChatGPT.

We selected three representative datasets: PestObserver-France, Natural-Hazards-Type, and Amazon-Food-Comments, and collected a set of response examples based on these datasets. Firstly, we calculated the proportion of responses that can be directly parsed using the Rule-based matching strategy, as shown in the "extra rate" column in Table 6. From this column, it can be observed that over 98% of the responses fall into the category of simple answers, which can be

Table 6

The several collected typical cases of exceptional responses from the ChatGPT Q&A step on three agricultural text classification datasets.

Dataset	Categories	Extra rate	Responses
PestObserver-France	Bioagressor,Disease,Others	0.422%	Bioagressor (pesticides/fungicides ... Disease, (Bauhinia blight or Zinnia blight ...
Natural-Hazards-Type	Hurricane, Wildfires, ...	2.533%	Hurricane (Note: Typhoon and Hurricane are ... None of the above categories. The sentence ...
Amazon-Food-Comments	Negative,Positive,Neutral	1.998%	positive (although the word “unfortunately” ... mixed/neutral

Template Classify the following sentence into one of the given categories:
 [CATE] \n \n Sentence: [SENT] \n Category:

[SENT]: *maladies tavelures; bouillies bordelaise et bourguignonne, oxychlorure de cuivre, sulfate basique,...*

[CATE]: ['Bioagressor', 'Disease', 'Others']

ChatGPT Response:

Answer1:
Bioagressor.

Answer2:
Bioagressor (pesticides/fungicides).

[SENT]: *Typhoon Fitow Floods a Car Collection in China... More here:...*

[CATE]: ['Hurricane', 'Wildfires', 'Blizzard', 'Floods', 'Tornado']

ChatGPT Response:

Answer1:
Hurricane/Typhoon (Note: Both Typhoon and Hurricane fall under the category of 'Tropical Cyclones').

Answer2:
The sentence "Typhoon Fitow Floods a Car Collection in China... More here:..." <http://fb.me/11rOnLYfc> belongs to the category 'Floods'.

[SENT]: 植保无人机开展小麦田化学除草对农药有哪些要求? ...

[CATE]: [农业经济, 农业工程, 水产渔业, 养殖技术, 林业, 园艺, 农作物]

ChatGPT Response:

Answer1:
The sentence belongs to the category '农业工程' (Agricultural Engineering).

Answer2:
The sentence belongs to '农业工程'. Additionally, the mention of wheat as a crop and its significant yield connects it to the category of crops '农作物'.

Fig. 12. The intermediate responses of ChatGPT on three cross-linguistic agricultural classification datasets showcase its remarkable proficiency in delivering precise answers and elaborating on the underlying reasoning behind them.

successfully parsed using the *rule-based matching strategy*. From the statistics, approximately 1% to 3% of the responses in these three datasets require further parsing, which can be achieved through the application of our designed *similarity-based matching strategy*, allowing for complete resolution. For example, observing the “response” column, in the third case where ChatGPT’s answer is Hurricane (Note: Typhoon and Hurricane are . . . , this statement can be successfully categorized as belonging to the “hurricane” category within natural disasters through our *similarity-based matching strategy*. Similarly, in the sixth case, based on the *similarity-based matching strategy*, the ChatGPT’s answer “mixed/neutral” can be successfully classified as “neutral” instead of “negative” or “positive” because it has the shortest character edit distance compared to the “neutral” category for that response.

Overall, although ChatGPT typically generates responses based on the manual prompts we have designed, ignoring differences in word capitalization, there are still a few cases where it produces responses that cannot be confidently assigned to a predetermined category. Coincidentally, the similarity matching strategy effectively addresses this limitation. By integrating rule-based matching and similarity-based matching strategies and employing an answer alignment approach, it is possible to categorize and map the responses of ChatGPT into different categories, ensuring a high level of accuracy.

6.2. Prominent comprehension capability of ChatGPT

In this comprehensive analysis, we delve into ChatGPT’s exceptional ability to understand nuanced semantics and accurately identify errors.

We present a collection of observed intermediate ChatGPT answers, as illustrated in Fig. 12, obtained from three utilized datasets. During API calls,¹⁷ we introduced randomness in generating different responses with ChatGPT by adjusting the probability and assigned two distinct responses for each query.

Using ChatGPT, we successfully classify the document segment from the PestObserver-France dataset into predefined categories, as depicted on the left side of Fig. 12. ChatGPT not only accurately identifies it as belonging to the “bioagressor” category but also showcases its ability to discern more nuanced features, such as “pesticides/fungicides”.

Furthermore, ChatGPT demonstrates exceptional semantic comprehension skills, regardless of the question format — be it English, French, or a combination of multiple languages. This is evident in its precise categorization of the document. Even in a bilingual example involving both Chinese and English (the third example), ChatGPT showcases a remarkable level of understanding and delivers fluent responses. These instances further highlight the impressive and prominent semantic comprehension capabilities of ChatGPT. We summarize the strengths and characteristics of ChatGPT in text classification tasks as follows:

- Accurate cross-lingual classification: ChatGPT exhibits excellent cross-lingual semantic understanding and can fluently answer

¹⁷ ChatGPT official API: <https://api.openai.com/v1/engines/davinci-codex/completions> [Accessed on 2023.06].

questions. Whether it is English, French, or a combination of multiple languages, ChatGPT can accurately classify documents into predefined categories;

- Fine-grained semantic understanding: ChatGPT demonstrates sensitivity to subtle semantic differences, accurately identifying the category of the agricultural document and discovering more nuanced features, such as “pesticides/fungicides”;
- Multiple random responses generation: By introducing randomness, ChatGPT can generate different responses for each query, adding diversity and allowing ChatGPT to exhibit varied responses when answering agricultural questions.

In conclusion, ChatGPT excels in agricultural text classification tasks with its exceptional accuracy in classification, sensitivity to semantic understanding, outstanding cross-lingual semantic comprehension, and the ability to generate multiple random responses.

7. Conclusion and outlook

Agricultural text classification, which serves as the basis for organizing various types of documents, is a crucial step towards managing massive and ever-increasing agricultural information. Existing PLM-based classification models face challenges such as dependency on annotated corpora, limited cross-linguistic transferability, and complex deployment. The emergence of ChatGPT has provided a solution to these issues. Nevertheless, there is a lack of systematic exploration of the benefits that ChatGPT brings to sustainable agricultural information management, especially in agricultural text classification. To this end, we explored ChatGPT’s potential in agricultural text classification in this work and proposed ChatAgri, a novel framework. ChatAgri is the first qualitative analysis of text classification on ChatGPT in the agricultural domain. Our experiments compared ChatAgri with various baselines, including traditional ML methods, PLM-based fine-tuning, and prompt learning. We tested datasets in multiple languages and developed prompt generation strategies to enhance ChatGPT’s generation capabilities. Furthermore, we evaluated GPT-4’s performance through comparative experiments. Overall, our results demonstrate the superiority of ChatGPT in agricultural text classification. This empirical exploration has opened new milestones for ChatGPT-based agricultural information management. We aim to propose more applications in sustainable agricultural development, promoting digital transformation and efficiency. For instance, ChatGPT aids smart agriculture, enhancing crop management and production quality. Overall, we hope this work showcases LLM’s potential in agriculture, leveraging existing knowledge databases for exciting AI advancements.

CRediT authorship contribution statement

Biao Zhao: Conceptualization, Investigation, Methodology, Project administration, Formal analysis, Writing – original draft, Writing – review & editing, Validation, Funding acquisition. **Weiqliang Jin:** Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing, Formal analysis, Validation, Visualization. **Javier Del Ser:** Writing – review & editing, Supervision, Funding acquisition. **Guang Yang:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments, corrections, and recommendations, which significantly improved the quality of the paper. G.Y. was supported in part by the ERC IMI, UK (101005122), the H2020 (952172), the MRC, UK (MC/PC/21013), the Royal Society, UK (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, UK, the SABER project supported by Boehringer Ingelheim Ltd, UK, and the UKRI Future Leaders Fellowship, UK (MR/V023799/1). J.D.S. also acknowledged support from the Spanish *Centro para el Desarrollo Tecnológico Industrial* (CDTI), Spain through the AI4ES project, and the Department of Education of the Basque Government (*Eusko Jaurlaritz*), Spain via the Consolidated Research Group MATHMODE (IT1456-22). B.Z. and W.J. were supported in part by the *Natural Science Basis Research Plan in Shaanxi Province of China* (Project Code: 2021JQ-061). Both the first two authors, B.Z. and W.J., made equal contributions to this work.

References

- [1] S. Jiang, R. Angarita, S. Cormier, F. Rousseaux, Fine-tuning BERT-based models for plant health bulletin classification, 2021, [arXiv:2102.00838](https://arxiv.org/abs/2102.00838).
- [2] Y. Xiao, Q. Du, Statistical age-of-information optimization for status update over multi-state fading channels, 2023, [arXiv:2303.11153](https://arxiv.org/abs/2303.11153).
- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, <https://doi.org/10.18653/v1/N19-1423>.
- [4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880, <https://doi.org/10.18653/v1/2020.acl-main.703>.
- [5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (1) (2020).
- [6] J.-L. Xu, Y.-L. Hsu, Analysis of agricultural exports based on deep learning and text mining, *J. Supercomput.* 78 (8) (2022) 10876–10892, <https://doi.org/10.1007/s11227-021-04238-w>.
- [7] Y. Cao, Z. Sun, L. Li, W. Mo, A study of sentiment analysis algorithms for agricultural product reviews based on improved BERT model, *Symmetry* 14 (8) (2022) <https://doi.org/10.3390/sym14081604>.
- [8] F. Hua Leong, C. Farn Haur, Deep learning-based text recognition of agricultural regulatory document, in: C. Bădică, J. Treur, D. Benslimane, B. Hnatkowska, M. Krótkiewicz (Eds.), *Advances in Computational Collective Intelligence*, Springer International Publishing, Cham, 2022, pp. 223–234.
- [9] W. Jin, H. Yu, X. Luo, CvT-ASDD: Convolutional vision-transformer based attentive single shot MultiBox detector, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence, (ICTAI), 2021, pp. 736–744, <https://doi.org/10.1109/ICTAI52525.2021.00117>.
- [10] H. Liu, N. Wang, X. Li, C. Xu, Y. Li, BFF R-CNN: Balanced feature fusion for object detection, *IEICE Trans. Inf. Syst.* 105 (8) (2022) 1472–1480.
- [11] N. Xia, H. Yu, Y. Wang, J. Xuan, X. Luo, DAFS: A domain aware few shot generative model for event detection, *Mach. Learn.* 112 (3) (2023) 1011–1031, <https://doi.org/10.1007/s10994-022-06198-5>.
- [12] J. Gao, H. Yu, S. Zhang, Joint event causality extraction using dual-channel enhanced neural network, *Knowl.-Based Syst.* 258 (2022) 109935, <https://doi.org/10.1016/j.knsys.2022.109935>.
- [13] T. Eloundou, S. Manning, P. Mishkin, D. Rock, GPTs are GPTs: An early look at the labor market impact potential of large language models, 2023, [arXiv:2303.10130](https://arxiv.org/abs/2303.10130).
- [14] OpenAI, GPT-4 technical report, 2023, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [15] L. Qing, T. Josh, E.Z. Michael, P. Janardhana, N. Chuang, et al., Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential, *Vis. Comput. Ind. Biomed. Art.* 6(1) (9) (2023) 10965–10973, <https://doi.org/10.1186/s42492-023-00136-5>.
- [16] T. Susnjak, Applying BERT and ChatGPT for sentiment analysis of lyme disease in scientific literature, 2023, [arXiv:2302.06474](https://arxiv.org/abs/2302.06474).
- [17] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, B. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner,

- S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 182–207.
- [18] W. Jin, B. Zhao, H. Yu, X. Tao, R. Yin, G. Liu, Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning, *Data Min. Knowl. Discov.* (2022) <http://dx.doi.org/10.1007/s10618-022-00891-8>.
- [19] W. Jin, B. Zhao, L. Zhang, C. Liu, H. Yu, Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis, *Inf. Process. Manage.* 60 (3) (2023) 103260, <http://dx.doi.org/10.1016/j.ipm.2022.103260>.
- [20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM Comput. Surv.* 55 (9) (2023) <http://dx.doi.org/10.1145/3560815>.
- [21] Y. Shen, X. Ma, Z. Tan, S. Zhang, W. Wang, W. Lu, Locate and label: A two-stage identifier for nested named entity recognition, 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)*, Association for Computational Linguistics, Online, 2021, pp. 2782–2794, <http://dx.doi.org/10.18653/v1/2021.acl-long.216>.
- [22] Y. Shen, X. Wang, Z. Tan, G. Xu, P. Xie, F. Huang, W. Lu, Y. Zhuang, Parallel instance query network for named entity recognition, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 947–961, <http://dx.doi.org/10.18653/v1/2022.acl-long.67>.
- [23] M.U. Haque, I. Dharmadasa, Z.T. Sworna, R.N. Rajapakse, H. Ahmad, “I think this is the most disruptive technology”: Exploring sentiments of ChatGPT early adopters using Twitter data, 2022, [arXiv:2212.05856](https://arxiv.org/abs/2212.05856).
- [24] J. Gao, H. Zhao, C. Yu, R. Xu, Exploring the feasibility of ChatGPT for event extraction, 2023, [arXiv:2303.03836](https://arxiv.org/abs/2303.03836).
- [25] Q. Zhong, L. Ding, J. Liu, B. Du, D. Tao, Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT, 2023, [arXiv:2302.10198](https://arxiv.org/abs/2302.10198).
- [26] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, W. Han, Zero-shot information extraction via chatting with ChatGPT, 2023, [arXiv:2302.10205](https://arxiv.org/abs/2302.10205).
- [27] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, GPT understands, too, 2021, [arXiv:2103.10385](https://arxiv.org/abs/2103.10385).
- [28] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, J. Tang, P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 61–68, <http://dx.doi.org/10.18653/v1/2022.acl-short.8>.
- [29] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, et al., A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT, 2023, [arXiv:2302.09419](https://arxiv.org/abs/2302.09419).
- [30] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, J. Zhou, Is ChatGPT a good NLG evaluator? A preliminary study, 2023, [arXiv:2303.04048](https://arxiv.org/abs/2303.04048).
- [31] N. Azeez, I. Al-Taie, W. Yahya, A. Basbrain, A. Clark, Regional agricultural land texture classification based on GLCMs, svm and decision tree induction techniques, in: *2018 10th Computer Science and Electronic Engineering (CEECE)*, 2018, pp. 131–135, <http://dx.doi.org/10.1109/CEECE.2018.8674193>.
- [32] Y. Li, S. Zhang, C. Lai, Agricultural text classification method based on dynamic fusion of multiple features, *IEEE Access* 11 (2023) 27034–27042, <http://dx.doi.org/10.1109/ACCESS.2023.3253386>.
- [33] J. Dunnmon, S. Ganguli, D. Hau, B. Husic, Predicting US state-level agricultural sentiment as a measure of food security with tweets from farming communities, 2019, [arXiv:1902.07087](https://arxiv.org/abs/1902.07087).
- [34] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, *OpenAI Blog* (2018).
- [35] S. Yunlai, C. Yunpeng, D. Zhigang, A classification method of agricultural news text based on BERT and deep active learning, *J. Libr. Inf. Sci. Agric.* 34 (8) (2022) 19, <http://dx.doi.org/10.13998/j.cnki.issn1002-1248.22-0172>.
- [36] J.-L. Xu, Y.-L. Hsu, Analysis of agricultural exports based on deep learning and text mining, *J. Supercomput.* 78 (8) (2022) 10876–10892, <http://dx.doi.org/10.1007/s11227-021-04238-w>.
- [37] C. Edio da, T. Handayani, D. Supeno, Text mining for pest and disease identification on rice farming with interactive text messaging, *Int. J. Electr. Comput. Eng.* 8 (3) (2020) 1671–1683, <http://dx.doi.org/10.11591/ijece.v8i3.pp1671-1683>.
- [38] R. Alec, W. Jeffrey, C. Rewon, L. David, A. Dario, S. Ilya, Language models are unsupervised multitask learners, *OpenAI Blog* (2019).
- [39] W. Jin, B. Zhao, C. Liu, Fintech key-phrase: A new Chinese financial high-tech dataset accelerating expression-level information retrieval, in: X. Wang, M.L. Sapino, W.-S. Han, A. El Abbadi, G. Dobbie, Z. Feng, Y. Shao, H. Yin (Eds.), *Database Systems for Advanced Applications, Springer Nature Switzerland, Cham*, 2023, pp. 425–440.
- [40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.
- [41] Z. Nanyang, L. Xu, L. Ziqian, H. Kai, W. Yingkuan, et al., Deep learning for smart agriculture: Concepts, tools, applications, and opportunities, *Int. J. Agric. Biol. Eng.* 11 (4) (2018) 32–44, <http://dx.doi.org/10.25165/j.ijabe.20181104.4475>.
- [42] S. Hu, N. Ding, H. Wang, Z. Liu, J. Wang, J. Li, W. Wu, M. Sun, Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2225–2240, <http://dx.doi.org/10.18653/v1/2022.acl-long.158>.
- [43] J. Fang, Z. Wang, W. Liu, S. Lauria, N. Zeng, C. Prieto, F. Siktström, X. Liu, A new particle swarm optimization algorithm for outlier detection: Industrial data clustering in wire arc additive manufacturing, *IEEE Trans. Autom. Sci. Eng.* (2022) 1–14, <http://dx.doi.org/10.1109/TASE.2022.3230080>.
- [44] Y. Zhang, P. Wu, H. Li, Y. Liu, F.E. Alsaadi, N. Zeng, DPF-S2S: A novel dual-pathway-fusion-based sequence-to-sequence text recognition model, *Neurocomputing* 523 (2023) 182–190, <http://dx.doi.org/10.1016/j.neucom.2022.12.034>.
- [45] P. Wu, Z. Wang, B. Zheng, H. Li, F.E. Alsaadi, N. Zeng, AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion, *Comput. Biol. Med.* 152 (2023) 106457, <http://dx.doi.org/10.1016/j.combiomed.2022.106457>.
- [46] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q.V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, 2023, [arXiv:2302.04023](https://arxiv.org/abs/2302.04023).
- [47] W. Jiao, W. Wang, J. tse Huang, X. Wang, Z. Tu, Is ChatGPT a good translator? Yes with GPT-4 as the engine, 2023, [arXiv:2301.08745](https://arxiv.org/abs/2301.08745).
- [48] Y. Wang, Q. Yao, J.T. Kwok, L.M. Ni, Generalizing from a few examples: A survey on few-shot learning, *ACM Comput. Surv.* 53 (3) (2020) <http://dx.doi.org/10.1145/3386252>.
- [49] L. Meng, Z.S. Dong, Natural hazards Twitter dataset, 2020, [arXiv:2004.14456](https://arxiv.org/abs/2004.14456).
- [50] Y. Kim, Convolutional neural networks for sentence classification, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, <http://dx.doi.org/10.3115/v1/D14-1181>.
- [51] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI '16*, AAAI Press, 2016, pp. 2873–2879.
- [52] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, M. Sun, OpenPrompt: An open-source framework for prompt-learning, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 105–113, <http://dx.doi.org/10.18653/v1/2022.acl-demo.10>.
- [53] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543, <http://dx.doi.org/10.3115/v1/D14-1162>.

Biao Zhao was born in 1986 in Xi'an, China. He obtained his Ph.D. degree in electronics and communication engineering from Politecnico di Torino in 2015 with the thesis View point robust visual search technique. He obtained his M.Sc. degree in Communication Engineering also from Politecnico di Torino in 2011. His Master thesis was about the image compression for solar probe which was collaborated with METIS in Italy. He obtained this Bachelor's degree in Information engineering in 2009 from Xi'an Jiaotong University in China. He is now an Assistant Professor at Xi'an Jiaotong University.

Weiqliang Jin is a Ph.D. student at Xi'an Jiaotong University in China.

Javier Del Ser was born in Barakaldo (Spain) in 1979. He joined the Faculty of Engineering (ETSI, www.ingeniaritza-bilbao.ehu.es) of the University of the Basque Country (Spain) to study Electrical Engineering, obtaining his combined B.S. and M.S. degree in May 2003. After finishing this degree, he became a recipient of the Fundacion de Centros Tecnologicos Inaki Goenaga doctoral grant. He received his first Ph.D. in Telecommunication Engineering (Cum Laude) from the University of Navarra, Spain, in 2006, and a second Ph.D. in Computational Intelligence (Summa Cum Laude, Extraordinary Prize) from the University of Alcala, Spain, in 2013. Currently he is a principal researcher in data analytics and optimization at TECNALIA (Spain).

Guang Yang is an Associate Professor (Senior Lecturer) at Bioengineering Department and Imperial-X, Imperial College London. Dr Guang Yang is a UKRI Future Leaders Fellow. He is also an Honorary Senior Lecturer in the School of Biomedical Engineering & Imaging Sciences at King's College London. His research group is interested in developing novel and translational techniques for imaging and biomedical data analysis. His group focuses on the research and development of data-driven fast imaging, data harmonization, image segmentation, image synthesis, federated learning, explainable AI etc. He is currently working on a wide range of clinical applications in cardiovascular disease, lung disease and oncology. Read more information about Yang's Lab at: <https://www.yanglab.fyi/>.