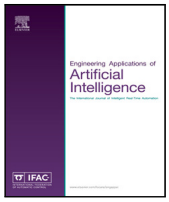
Engineering Applications of Artificial Intelligence 132 (2024) 107907

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



AGCVT-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning

Xu Gu [a], Xiaoliang Chen [a,b,*], Peng Lu [b], Zonggen Li [c], Yajun Du [a], Xianyong Li [a]

[a] School of Computer and Software Engineering, Xihua University, Chengdu 610039, China
[b] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C3J7, Canada
[c] Department of Computer Science, LMU Munich, Munich 80539, Germany

ARTICLE INFO

Keywords:
Large language models
Prompt learning
Sentiment classification
Chain of thought

ABSTRACT

Large language models (LLMs) have revolutionized natural language processing, but they require significant data and hardware resources. Prompt learning offers a solution by enabling a single model for multiple downstream tasks. However, current prompt learning methods rely on costly prompt templates for training. This is a challenge for tasks like sentiment classification, where high-quality templates are hard to create and pseudo-token composed templates can be expensive to train. Recent studies on the chain of thought (COT) have shown that enhancing the presentation of certain aspects of the reasoning process can improve the performance of LLMs. With this in mind, this research introduces the auto-generated COT and verbalizer templates (AGCVT-Prompt) technique, which clusters unlabeled texts according to their identified topic and sentiment. Subsequently, it generates dual verbalizers and formulates both topic and sentiment prompt templates, utilizing the categories discerned within the text and verbalizers. This method significantly improves the transparency and interpretability of the model's decision-making processes.

The AGCVT-Prompt technique was evaluated against conventional prompt learning and advanced sentiment classification methods, using state-of-the-art LLMs on both Chinese and English datasets. The results showed superior performance in all evaluations. Specifically, the AGCVT-Prompt method outperformed previous prompt learning techniques in few-shot learning scenarios, providing higher zero-shot and few-shot learning capabilities. Additionally, AGCVT-Prompt was utilized to analyze network comments about Corona Virus Disease 2019, providing valuable insights. These findings indicate that AGCVT-Prompt is a promising alternative for sentiment classification tasks, particularly in situations where labeled data is scarce.

1. Introduction

The rise of social media has significantly changed the way citizens communicate with each other. Online social networks now play a vital role in how people acquire information and express their opinions. However, this digital platform also poses some challenges, especially when it comes to sensitive topics. Extreme or radical comments can have a significant impact on how the public perceives and feels about certain issues. For instance, a critical post about Japan's decision to release nuclear-contaminated water into the ocean may go unnoticed. But during the COVID-19 pandemic, a post that expresses strong sentiments and denounces governmental inaction, even if it is not entirely accurate, can cause anxiety, undermine governmental credibility, and disrupt social harmony. Therefore, it is crucial to identify key sentimental indicators in social media discourse so that the government can manage and steer online public opinion.

Numerous deep learning models, including Long Short-Term Memory (LSTM) (Greff et al., 2015) and Gated Recurrent Units (GRU) (Cho et al., 2014), have been developed for sentiment recognition research in the field of Natural Language Processing (NLP). Predominantly, Language Models (LMs) and large LMs (LLMs) employing Transformer architectures leverage extensive textual datasets, facilitating their comprehension of linguistic patterns and structures. This capability enables them to execute intricate tasks. The swift advancement in high-performance Graphics Processing Units (GPUs) empowers researchers to configure more parameters for training LLMs. In recent developments, multi-turn dialogue generation models, such as ChatGPT (Brown et al., 2020), LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023) and Falcon (Penedo et al., 2023), are notable for using an augmented parameter set, reflecting the ongoing evolution in this domain. Fig. 1 depicts the evolution of the number of parameters used in NLP models over time.

* Corresponding author at: Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C3J7, Canada.
E-mail address: chexiaol@iro.umontreal.ca (X. Chen).

https://doi.org/10.1016/j.engappai.2024.107907
Received 9 August 2023; Received in revised form 14 December 2023; Accepted 12 January 2024
Available online 19 January 2024

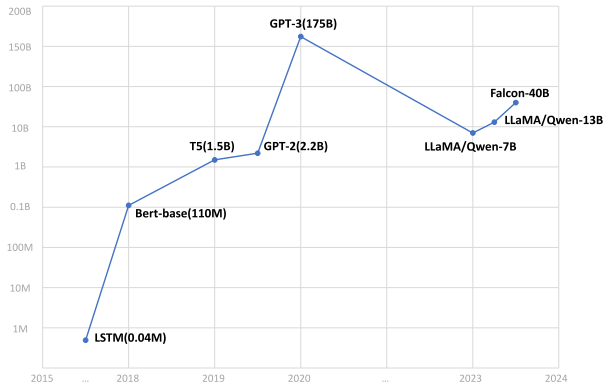


Fig. 1. Specification trends for mainstream NLP models from 2015 to 2023.

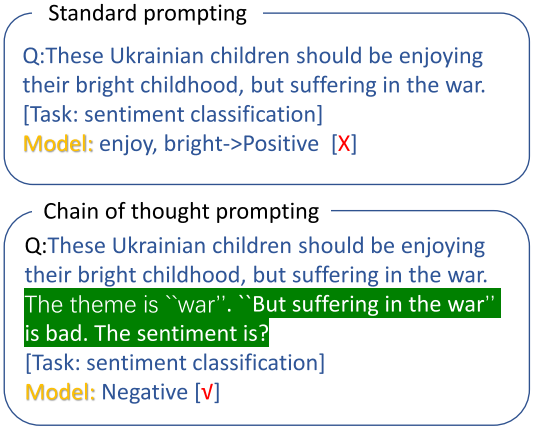


Fig. 2. Proposed AGCVT-Prompt method uses chains of thought to help LLMs reason about sentiment classification tasks.

Generative LLMs have demonstrated remarkable capabilities, yet their training process is resource-intensive, necessitating substantial refined data and sophisticated hardware. Further refinement of these models often involves reinforcement learning with human feedback to optimize interaction performance. Consequently, despite the availability of open-source generative models (ranging from 3B to 70B parameters), their usage remains limited. LLMs with significantly fewer parameters necessitate adaptation for various downstream tasks using conventional fine-tuning methods. Tailoring these models for specific tasks frequently demands considerable resources, rendering it impractical for nascent NLP applications. Furthermore, developing dedicated LLMs for each specific task is often unfeasible, as only a select number of teams possess the requisite resources for such an endeavour.

Prompt learning is an efficient and cost-effective technique that has several advantages over traditional deep learning methods. Unlike fine-tuning, prompt learning requires only a small amount of labeled data. Additionally, using natural language prompts allows researchers to guide the model to produce specific outputs while still allowing it to generate its outputs. This flexibility makes prompt learning highly adaptable to different research targets. Moreover, the use of natural language prompts enhances the interpretability of the model's outputs, allowing researchers to better understand how the model generates its outputs.

However, prompt learning also presents some challenges that need to be addressed. One of the primary issues is the complexity of creating effective prompts that can generate high-quality outputs. Inappropriate prompts can lead to biased outputs and weaken the performance of LLMs. Moreover, current prompt learning methods often focus on manually constructed or auto-generated prompt templates. The impact of verbalizers on LLMs has also not been fully explored. For researchers, building appropriate prompt templates and verbalizers for complex classification tasks can be a challenging task that requires to be tackled.

Our approach suggests using pre-generated templates in combination with soft templates to automatically construct COT prompts, with verbalizers generated through a generative model as shown in Fig. 2. Researchers can choose between manual or automatic prompt and verbalizer construction to facilitate the prompt learning process.

Sentiment classification research is of utmost importance in the management of public opinion and opinion evolution, especially in the current era of international turbulence and post-pandemic uncertainties. In this regard, we propose an auto-generated COT and verbalizer templates method, AGCVT-Prompt, to perform sentiment classification on two datasets of network comments in Chinese and English languages. Our experiments involve T5 (Raffel et al., 2019) and BERT (Devlin et al., 2018) models on six datasets and demonstrate superior performance compared to existing prompt learning and advanced sentiment classification techniques. We present a detailed analysis of the performance of our proposed AGCVT-Prompt method.

The main contributions of this study are summarized as follows:

- AGCVT-Prompt innovatively integrates the HDBSCAN clustering algorithm with the T5 model, automating thematic subject extraction and verbalizer development. This approach eliminates manual template construction, enhancing the reasoning capabilities of generative models.
- The proposed AGCVT-Prompt utilizes automated COT technology to significantly improve prompt learning templates' efficiency. It overcomes the limitations of manual COT prompts, ensuring higher effectiveness in diverse applications.
- AGCVT-Prompt exhibits exceptional adaptability in scenarios with sparse data, such as in public health contexts during crises like the COVID-19 pandemic. It efficiently facilitates both binary and multi-class sentiment classification with minimal data requirements, demonstrating strong generalization and scalability.
- AGCVT-Prompt is among the first methods to integrate COT prompt technology in sentiment classification, showcasing superior performance. It outperforms advanced techniques across multiple datasets and is compatible with lightweight LLMs, marking a significant advancement in the field.
- Compared to existing baselines, AGCVT-Prompt shows superior performance in few-shot and zero-shot learning scenarios. This method allows for cost-effective experimentation, with results underscoring its potential in practical applications.
- In the complex domain of public health data analysis, AGCVT-Prompt substantially enhances model explainability. It refines sentiment classification precision and elucidates the decision-making processes, which is crucial for understanding model outputs.

The following paper is structured as follows. Section 2 presents a comprehensive review of recent relevant research. Section 3 introduces the proposed AGCVT-Prompt model. The experimental settings are outlined in Section 4, and the findings of the experiments are analyzed in Section 5. Finally, the last section provides a summary of the research presented in this paper. All the datasets and codes are publicly available online.[1]

2. Related work

In this section, we will provide an overview of recent developments and advancements in the field of NLP. Specifically, we will discuss progress made in three areas: large language models, prompt learning, and chain of thought.

[1] https://github.com/gooSAMA/AGCVT.

2.1. Large language models

After the emergence of the generative multi-round question-answering model, researchers are accustomed to calling architectures with relatively small specifications, such as Bert (Devlin et al., 2018) and Bart (Lewis et al., 2020), the language models (LMs), and models such as ChatGPT (Brown et al., 2020) and Qwen (Bai et al., 2023), large language models (LLMs). Despite demonstrating great potential in various NLP applications, LLMs continue to face major challenges such as poor robustness and weak interpretability. Generative LLMs are characterized by their propensity to produce variable responses. Identical prompts elicit disparate outputs influenced by contextual elements from previous dialogues. Slight semantic variations in prompts can precipitate markedly divergent responses. In contrast, MLM frameworks consistently generate a singular output in response to a given prompt, exhibiting enhanced robustness achieved through optimization processes. Furthermore, a notable challenge in generative models is their lack of explanatory sufficiency, which impedes elucidating interpretable reasoning processes and mechanisms. The step-wise logical reasoning facilitated by AGCVT-Prompt not only aids in refining model judgments but also offers clear explanatory pathways, significantly augmenting the model's interpretability.

The rise of LMs and LLMs can be attributed to the advancements in Transformer technology (Vaswani et al., 2017). The seq2seq architecture of Transformer is composed of an encoder and decoder, utilizing attention mechanisms to capture long-term semantic dependencies in textual data. This development has greatly enhanced the performance of various NLP tasks and enabled transfer learning through the creation of pre-trained language models.

One pre-trained language model that has gained widespread popularity in the NLP community is BERT (Devlin et al., 2018), which was introduced by Devlin et al. BERT utilizes the masked language model (MLM) approach to create deep bidirectional language representations by employing bidirectional Transformers. Additionally, BERT uses fine-tuning techniques to specialize in specific tasks. Other models based on BERT have been introduced by adjusting and reducing their parameters, such as RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019).

In addition to BERT, there are other popular pre-trained language models based on Transformer technology, such as the Generative Pre-trained Transformer (GPT) series, comprising GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023). These models also use masked multi-head attention and left-to-right generative mechanisms. They are pre-trained on vast amounts of data in an unsupervised manner and then fine-tuned on smaller supervised datasets for specific tasks. However, despite GPT-3's enormous size of 175 billion parameters, the quality of the generated text is not always satisfactory.

LMs have not yet received significant attention from researchers, as mainstream research still heavily relies on BERT fine-tuning. However, GPT-3 has introduced the concept of prompts for LLMs. For instance, when given the input "This movie is really good, my mood is _". GPT-3 generates the output "positive". Both BERT and GPT-3 demonstrate that large models still suffer from weak interpretability, poor robustness, and limited reasoning capabilities common in deep learning models. Raffel et al. introduced the T5 model, an LLM with a powerful capacity to accomplish specific downstream tasks by inputting different prompt words manually, without requiring fine-tuning (Raffel et al., 2019). T5 can convert all NLP tasks into a Text-to-Text format using the same model, loss function, training process, and decoding process to accomplish all NLP tasks.

The emergence of ChatGPT has highlighted the potential of large generative models, leading to a surge in the development and training of expansive models (Bai et al., 2023; Touvron et al., 2023; Penedo et al., 2023), mainly in the ranges of 7B, 13B, and even up to 70B. These advanced models can engage in multi-turn dialogues, follow instructions, and perform reasoning tasks. However, creating suitable templates for these massive models is crucial to understanding user needs and providing tailored explanations. Despite the significant progress in generative dialogue AI, there are still significant challenges that the field needs to overcome. The most significant challenge is training and operationalizing these large models. For example, a Geforce 4060 8G graphics card can infer the 7B models, but it lacks the capacity for training or fine-tuning. Training more extensive models requires graphics cards with enhanced memory and performance capabilities. Additionally, the quality of prompts is critical. The variation in text quality generated from similar prompts can be significant, requiring considerable effort to identify high-quality prompts. The proposed method provides a solution by autonomously generating COT prompts in a cost-effective experimental setting, enabling LLMs to understand and reason more effectively.

Recent advancements in sentiment classification have seen the emergence of methodologies utilizing LM based approaches. Kaur and Kaur (2023) introduced a method that synergizes BERT with Bidirectional Convolutional Neural Networks for identifying sentiments within the text. This Bert-BiCNN approach employs dual CNNs to extract sentimental features from texts efficiently, subsequently integrating these features into the BERT architecture for processing. Additionally, Talaat (2023) developed HybridBert, an innovative model that amalgamates various BERT configurations. When applied to sentiment classification tasks, this model selects an appropriate BERT architecture for processing, demonstrating enhanced efficacy compared to employing a singular BERT model.

2.2. Prompt learning

Prompt Learning is a semi-supervised training paradigm for language models that aims to bridge the gap between their deep-level semantic understanding and human cognitive levels by redefining input examples as closed phrases called "prompts". This approach requires less parameter training than fully fine-tuning the model and has paved the way for subsequent prompt methods to transform downstream NLP tasks. The Pattern Exploiting Training (PET) method, proposed by Schick and Schütze (2021), was the first paradigm of prompt learning. It defines hand-crafted prompts as patterns and the answer mapping space as a verbalizer. PET helps language models understand tasks by defining prompts and verbalizers such as "positive", "negative", and "neutral" for sentiment classification tasks and "sunny", "cloudy", and "rainy" for weather forecasting tasks. Although PET's method requires continuous experimentation and adjustment to achieve a relatively good template, it trains fewer parameters than fully fine-tuning the model and lays the foundation for subsequent prompt methods to be developed.

Gao et al. proposed the LM-BFF method to improve the manual templates used in PET (Gao et al., 2021). To achieve this goal, they introduced auto-generated prompts, which streamlined brute-force searches to determine the optimal working label word and a new decoding objective of generating templates using the generative T5 model. Gao et al. generated a large number of prompt words using T5, which they fine-tuned one by one to obtain templates suitable for model training. Similar to Gao's work, we also utilized T5 to generate prompt templates and selected the most appropriate one.

There are two approaches to generating manual hard templates: manual design based on experience and automated search. However, neither approach is always superior to the other, as manual design can be limited by individual expertise, while automated search has issues with readability and interpretability. In contrast, Liu et al. (2021) proposed the p-tuning method, which avoids manual design altogether by introducing several optimizable pseudo prompt tokens directly into the input. This approach avoids overfitting and requires a minimal number of parameters to optimize. Both PET and p-tuning methods significantly reduce the number of training parameters compared to traditional fine-tuning methods, as depicted in Fig. 3.

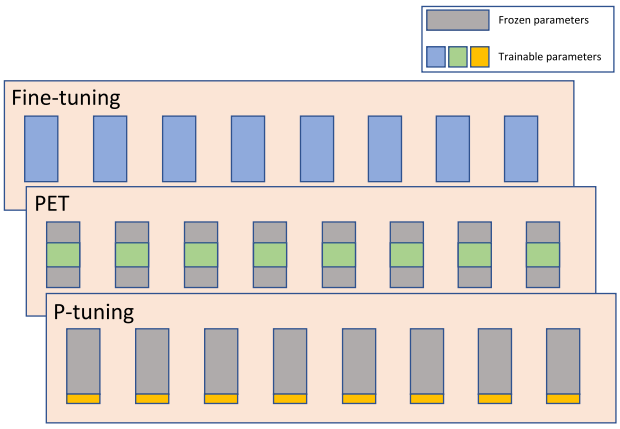


Fig. 3. Parameter scales of fine-tuning and prompt learning methods training on LLMs.

Li and Liang (2021) introduced the prefix-tuning approach to address the limited number of optimized parameters in updating prompt token embeddings in p-tuning. This method enables the optimization of a larger number of parameters, thereby enhancing the effectiveness of LLMs without imposing a significant computational burden.

Li and Liang (2021) proposed the prefix-tuning approach to overcome the limitation of a small number of optimized parameters in updating prompt token embeddings in p-tuning. This method enables the optimization of a larger number of parameters, thereby enhancing the effectiveness of LLMs without imposing a significant computational burden.

Soft prompt methods represent a class of NLP techniques that guide language models to perform a specific task without the need for hard-coded templates or prompts. These methods use probabilistic or continuous-valued signals to guide the model and can be automatically generated and optimized during model training, thereby reducing the need for manual intervention. However, the quality and effectiveness of soft prompts depend on the optimization algorithms and the quality of the training data. Despite their advantages, soft prompts rely heavily on the quality of the prompts used to guide the model. To address this, Zhou et al. (2022) proposed the APE method, which optimizes instructions by searching the candidate instruction pool suggested by the LLM and maximizing the selected scoring function. Inspired by the human-friendly approach of classical program synthesis and prompt engineering, the APE method achieved better or equivalent performance than human-generated instruction on multiple baselines.

Several methods have been proposed for optimizing pre-trained language models without the manual engineering of task-specific prompts or language models. Xu et al. (2023) proposed the CP-tuning framework, which utilizes a two-pair cost-sensitive contrastive learning process. This combines task-agnostic continuous prompt encoding techniques with fully trainable prompt parameters to improve model performance and enhance prompt invariance. Hu et al. (2022) developed a Knowledge-Prompted Tuning (KPT) technique, which integrates external knowledge into language models to improve and stabilize prompt tuning. KPT generates a set of label words for each label, covering different granularities and perspectives, and incorporates knowledge into the classification model through a multi-level Verbalizer. Li et al. (2021) proposed SentiPrompt, a unified framework that utilizes sentiment knowledge to enhance prompt tuning of language models. SentiPrompt constructs consistency and polarity judgment templates by extracting aspect-level text, opinions, and polarity triplets, and concatenates these elements into a single representation. Inspired by the excellent work of Li et al. the method proposed in this paper also constructs COT by extracting key information from the text. Finally, for the Implicit Discourse Relation Recognition (IDRR) task, Xiang et al. (2022) proposed two types of ConnPrompt templates: Insertion-Completion Prompts (ICP) and Prefix-Completion Prompts (PCP). ConnPrompt integrates multiple prompts to combine predictions from different prompt results and achieves better performance with less training data by constructing the answer space mapping for relation meanings based on hierarchical sense labels and implicit connectives.

2.3. Chain of thought

Researchers have recognized the limitations of continuous prompt learning, such as pseudo-resource conservation and instability. To overcome these issues, Wei et al. (2022) proposed the chain of thought (COT), a discrete prompt learning method that enhances the performance of generative language models by providing scattered language deductions. However, the effectiveness of LLM performance improvement is dependent on the quality of COT. Diao et al. (2023) proposed Active-Prompt, which uses task-specific example prompts with manually designed COT reasoning to adapt LLMs to different tasks. They also proposed four perspectives to evaluate prompt uncertainty and filter templates through uncertainty scores for reasoning tasks. Diao et al.'s innovative work has made a great contribution to the research of COT. Ho et al. (2022) proposed Fine-tune-CoT, which uses a large model as a teaching model to generate outputs through zero-shot thought chain reasoning. The generated reasoning samples, including questions and teacher model outputs, are used to fine-tune a smaller student model that can be applied to smaller pre-trained language models. Shi et al. (2023) found that adding an instruction to ignore irrelevant information significantly improves performance in benchmark tests. However, irrelevant information can still distract the model's attention and greatly reduce its performance. To address the limitations of manually generated text, we propose a method that utilizes automatically generated COT and prompt learning templates. The most suitable combination for LLMs is selected through performance evaluation. This approach avoids the shortcomings of manual text generation. Section 2 presents a comprehensive list of the latest advancements in prompt learning and rapid development of instruction learning. The default specification of Bert consists of 12 layers and 768 dimensions. The current prompt learning and COT methods rely on the manual creation of high-quality templates, which makes it crucial to develop automated ways to construct efficient prompts.

3. AGCVT-Prompt

The overall process of AGCVT-Prompt construction is shown in Fig. 4. The process starts with a collection of the original text, $S = \{S_1, S_2, \ldots, S_n\}$. This text is then transformed into features using the final hidden layer of the BERT framework. Next, the unsupervised HDBSCAN algorithm is applied to cluster the text. The results of this clustering are then fed into the T5 model, which generates thematic keywords based on the clustered topics. These keywords align the sentences within the original text, creating a topic-specific prompt template with the help of a prompt-based LLM. The next step is to use this prompt-based LM to perform a simple sentiment classification on the text, creating a sentiment-aware prompt template. These templates are then integrated into a soft prompt template of a specified length, along with prompts tailored to downstream tasks, to construct a comprehensive COT, which enables the model to reason in a structured, step-by-step manner (see Table 1).

In the scenario depicted in Fig. 2, the phrase "These Ukrainian children should be enjoying their bright childhood, but are suffering in the war" is subjected to processing via the AGCVT-Prompt model. This procedure autonomously generates a topic prompt template. The resultant components include: a thematic prompt template "The theme is 'war'", a sentiment prompt template "'But suffering in the war' is negative", and a downstream task template "What is the sentiment?". These elements collectively assist the model in comprehending and executing the classification task. It is noteworthy that soft tokens

Table 1
Comparison of excellent prompt learning methods in recent years.

Method	Description	Limitation
Fine-tuning (FT), 300M	FT is a common training approach that the entire model's parameters are updated during this process to maximize task-related performance.	FT requires a large amount of training data to update all parameters and is also computationally expensive.
Pattern Exploiting Training (PET), 0.07M	PET is a technique that uses pattern recognition to enhance model performance with limited data. It transforms the task into a fill-in-the-blank problem and allows the model to predict the missing pieces.	Model performance degradation may occur if the fill-in-the-blank problem is not translated appropriately, as it may lack context information.
P-tuning (PT),0.07M	PT includes an additional set of trainable pseudo-tokens that have no practical meaning to the input portion of the model. These tokens are optimized during training to help the model perform a specific task more accurately.	PT does not fully exploit the model's potential since only the prefixes of the input part are adjusted. This approach may not be sufficient for some complex tasks to achieve optimal performance.
Prefix-tuning, 0.7M	Prefix-tuning adds a prefix to each layer of Bert, increasing the number of trainable parameters to optimize for problems where PT does not fully exploit the model's potential.	The increase in the number of parameters takes longer. Low-quality prompts can seriously affect the model's quality.
Instruction learning/Chain of Thought (COT), 0.07M-0.7M	Prompt Learning and COT are both part of Instruction Learning. Instruction Learning is mainly used for generative LLMs. Constructing sound prompts maximizes the ability to stimulate LLMs' latent potential.	The current high-quality instruction sets are manually constructed using extensive data experiments, rendering them unsuitable for research scenarios with limited data and manufacturing implementation.

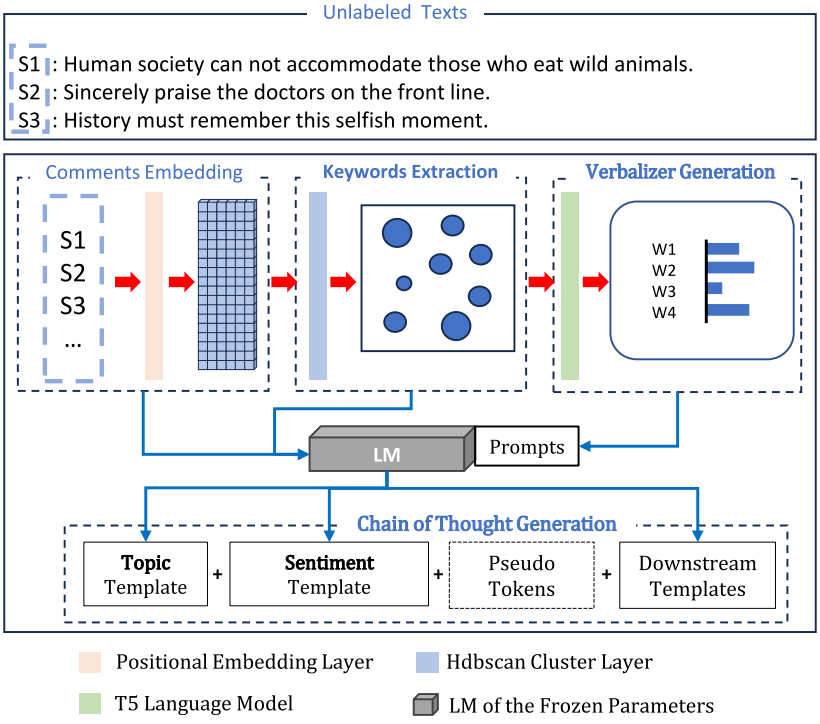


Fig. 4. The pipeline of constructing AGCVT-Prompt.

are not visually emphasized in this example, owing to the fact that these soft templates are essentially pseudo tokens, constituted by the specialized labels within the language model frameworks. Both topic templates and sentiment templates are generated based on T5 and then used to generate COT templates, so topic templates and sentiment templates are collectively referred to as pre-generated templates. The manual downstream templates are simply templates. Each module will be explained in greater detail in the following subsections.

3.1. Problem statement

Sentiment classification is an essential task in the field of NLP and machine learning. It aids in the comprehension and analysis of attitudes and opinions expressed in textual data, providing valuable insights for decision-making in various domains, such as marketing, politics, and customer service. The sentiment classification task involves assigning a sentiment label to a given text, indicating whether the underlying sentiment of the text is positive, negative, or neutral. This task can be formulated as a supervised learning problem, where a set of documents D and corresponding sentiment labels L are provided. The goal is to learn a mapping function f that can accurately assign a sentiment label l from L to each document d in D. To achieve this, a classifier is trained using a labeled training set to predict the sentiment labels of new, unlabeled documents.

3.2. Automatically generated verbalizer

The process of automatically generating verbalizers contains text feature representation, unsupervised topic clustering, and T5 model

generation of verbalizers. BERT can provide rich text feature representations that capture deep semantics and contextual information of texts.

First, input texts into the BERT model and use the output of the last hidden layer as feature representations. Let the text set be: $S = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes the ith text. Text feature extraction using pre-trained BERT model can be represented as:

$$x_i = f_{\text{BERT}}(s_i) \tag{1}$$

where $f_{\text{BERT}}(\cdot)$ is a mapping function that maps texts to fixed-length vectors. The high-dimensional features from BERT need proper dimension reduction for effective clustering. First, zero-center the text matrix X to obtain Z:

$$z_i = x_i - \overline{x} \tag{2}$$

where $\overline{x}$ is the mean vector of all x_i. Next, compute the covariance matrix $C \in R^{d\times d}$:

$$C = \frac{1}{n-1} Z Z^T \tag{3}$$

Then, compute the eigenvectors and eigenvalues of C. Let the eigenvector matrix be: $U = [u_1, u_2, \ldots, u_d]$, then the corresponding eigenvalues are:

$$\Sigma = [\sigma_1, \sigma_2, \ldots, \sigma_d] \tag{4}$$

Select the top d' principal components to construct the transformation matrix U': $U' = [u_1, u_2, \ldots, u_{d'}]$ The dimension-reduced projection is given by:

$$X' = ZU'^T \tag{5}$$

The processed text features are clustered using the HDBSCAN (Eldridge et al., 2015) method. Compared to K-means which can only find circular/spherical clusters, HDBSCAN can discover clusters of arbitrary shapes. Moreover, HDBSCAN is robust to outliers and noise. It marks those points as noise or boundary points instead of forcing them into clusters. The HDBSCAN clustering can be represented as:

$$C_i = f_{\text{HDBSCAN}}(x_i') \tag{6}$$

where $f_{\text{HDBSCAN}}(\cdot)$ is the HDBSCAN clustering function and C_i is the clustering output.

Finally, with the T5 model, a model that can be used for various text generation tasks, appropriate prompts are constructed based on the HDBSCAN clustering results and fed into the T5 model to generate descriptive texts for each cluster, which can be denoted as:

$$Verbalizer = \text{T5}_{p_i}(C_i) \tag{7}$$

where $\text{T5}_{p_i}(\cdot)$ denotes the T5 text generation operation with prompt p_i.

In particular, T5 can be used directly to generate a verbalizer for sentiment classification tasks based on prompts because binary and six-class sentiment categories can be exhaustive. The generation of a sentimental verbalizer using the T5 model is depicted in Fig. 5. Specifically, the process involves the use of $Prompt1$ to enable the T5 model to generate keywords w_i for each sentence of text S_i $(1 < i < m)$. Following this, $Prompt2$ is utilized to prompt the T5 model to group the keywords generated by $Prompt1$ into verbalizers. This process can be formalized as follows:

$$\begin{cases} [w_1, w_2, \ldots, w_m] = T5_{prompt1}([S_1, S_2, \ldots, S_m]) \\ [Label_1, Label_2, \ldots, Label_n] = T5_{prompt2}([w_1, w_2, \ldots, w_m]) \end{cases} \tag{8}$$

where $[S_1, S_2, \ldots, S_m] \subset D_{train}$.

3.3. Pre-generated prompt

Pre-generated prompts are a predetermined set of instructions or questions that are employed to elicit a response from a language model. These prompts are created based on specific criteria such as the nature of the task the model is intended to perform and are intended to guide the model toward more precise and relevant predictions. Inspired by Gao et al. (2021), the T5 model is applied to automatically generate templates $\mathcal{T}$ from a fixed set of label words $\mathcal{M}(\mathcal{Y})$. The generative model can construct the templates $\mathcal{T}$, without having to specify a predefined number of tokens for it, allowing simply taking input sentences from D_{train} for generating a set of different templates $\mathcal{T}$.

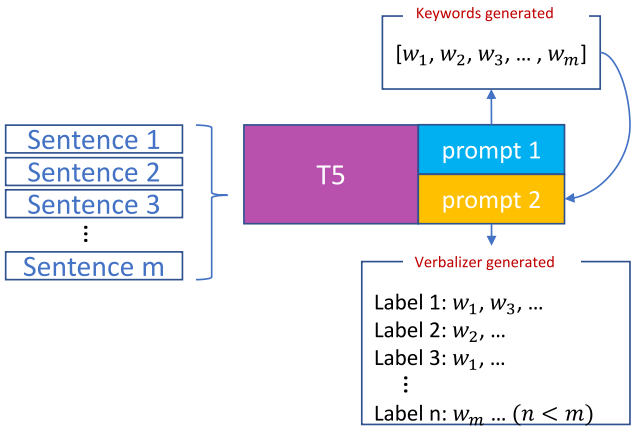


Fig. 5. The process of generating a verbalizer with T5 prompts.

Fig. 6 illustrates the pre-generated template method. Let $\mathcal{M} : \mathcal{Y} \mapsto \mathcal{V}$ be a mapping from the task label space to a single word in the vocabulary, where $\mathcal{V}$ denotes the verbalizer space of the prompt method. In Eq. (16), we employ the T5 model to complete the placeholders. Our objective is to generate an output that is effective for all the examples in $\mathcal{D}_{train}$, i.e., solving the decomposition formula expressed as follows:

$$\sum_{j=1}^{|\mathcal{T}|} \sum_{(x_i,y)\in D_{\text{train}}} \log P_{\text{T5}}\left(t_j \mid t_1, \ldots, t_{j-1}, \mathcal{T}\left(x_i, y\right)\right) \tag{9}$$

where P_{T5} represents the output probability distribution of T5, and $\left(t_1, \ldots, t_{|\mathcal{T}|}\right)$ are the template tokens. The default clustering results of HDBSCAN are used as a verbalizer and text labels before T5 input. However, some labels can also be manually corrected based on the clustering results, improving the accuracy of the generated COT. When constructing templates for different classification tasks, it is essential to provide specific downstream task prompt templates. For topic and sentiment classification, rigid templates can be formulated as "The topic of this text is [MASK]" and "The sentiment is [MASK]". Templates that closely align with the actual downstream task tend to yield more efficient results. Additionally, the definition of verbalizers is required. While AGCVT-Prompt enables automatic construction, manual creation should take into account the specific context of the downstream task. For instance, in topic classification, templates might include {"Sports": "football, basketball, swimming", "Entertainment": "games", "Movies": "cinema"}, while binary sentiment classification could involve {"Positive": "happy, joy, surprise", "Negative": "bad, anxiety, disgust"}. Similarly, for six-category sentiment classification, appropriate mappings might be {"Joy": "happy, joyful", "Sadness": "sad, sorrowful", "Fear": "fearful, scared", "Surprise": "surprised, amazed", "Disgust": "disgusted, repulsed"}. These tailored templates and verbalizers are pivotal in guiding the model towards accurate categorization in varied classification scenarios.

The validity of templates produced by the T5 model is assessed using D_{dev} and metrics are employed to filter them. T5 is specifically designed to complete missing portions (e.g., <X> or<Y>) in its input. To generate candidate verbalizers for each sentence, the input is structured as follows: "<Sentence> the keywords of this text is <X>, the sentiment is <Y>". T5 is proficient in generating phrases such as "<X> social_impact <Y> negative".

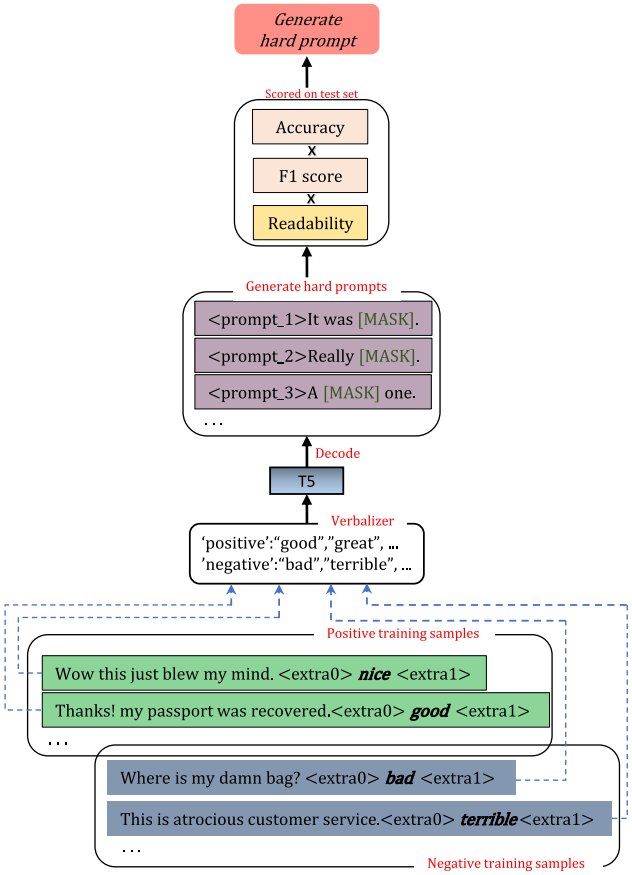


Fig. 6. The training process of pre-generating templates.

3.4. Automatically generated COT

Let the augmented input be represented as $\mathcal{T}_{\text{aug}}$, which is a concatenation of the original text $\mathcal{T}$, a set of reasoning steps $\{S_i\}_{i=1}^{n}$, and a sequence of soft prompts or pseudo-tokens $\{P_j\}_{j=1}^{m}$. Mathematically, it can be expressed as:

$$\mathcal{T}_{\text{aug}} = \mathcal{T} \oplus \bigoplus_{i=1}^{n} S_i \oplus \bigoplus_{j=1}^{m} P_j \tag{10}$$

where $\oplus$ denotes the concatenation operation. The COT aims to guide the BERT model $\mathcal{M}$ through a structured thought process, improving its ability to contextually analyze $\mathcal{T}$ and accurately perform sentiment classification. The reasoning steps $\{S_i\}$ provide explicit logical pathways, while the soft prompts $\{P_j\}$ introduce flexible, task-specific cues that assist in shaping the model's understanding and response. The output of the model, O, is the sentiment classification result derived from processing $\mathcal{T}_{\text{aug}}$. It is obtained as:

$$O = \mathcal{M}(\mathcal{T}_{\text{aug}}) \tag{11}$$

This approach leverages the inherent strengths of BERT in understanding natural language. At the same time, the COT structure introduces an additional layer of logical reasoning and contextual prompts, enhancing the model's performance on complex sentiment analysis tasks.

3.5. AGCVT-Prompt with LLMs

Fig. 7 presents the overall architecture of AGCVT-Prompt. The suggested prompt learning approach involves training a minimal number of additional parameters without fully fine-tuning the parameters pre-trained by LLMs. Given a sentence $s = [s_0, s_1, s_2, \ldots, s_n]$

Algorithm 1 Enhancing Sentiment Classification with AGCVT-Prompt in BERT

Input: Text sequence $\mathcal{T}$
Output: Sentiment classification result O.
$\mathcal{T}_{\text{aug}} \leftarrow \mathcal{T}$ \\ Augment Input $\mathcal{T}$.
for **do** $i = 1$ **to** n
  Append logical step S_i to $\mathcal{T}_{\text{aug}}$
end for
for **do** $j = 1$ **to** m
  Append soft prompt P_j to $\mathcal{T}_{\text{aug}}$
end for
return $\mathcal{T}_{\text{aug}}$ \\ Classify Sentiment $\mathcal{T}_{\text{aug}}$
$O \leftarrow \text{BERT}(\mathcal{T}_{\text{aug}})$
return Sentiment classification from O
$\mathcal{T}_{\text{aug}} \leftarrow \text{AugmentInput}(\mathcal{T})$
$O \leftarrow \text{Classify Sentiment}(\mathcal{T}_{\text{aug}})$

that consists of n words, it can be represented as embedded vectors $[e(s_0), e(s_1), e(s_2), \ldots, e(s_n)]$ after passing through the embedding layer. Typically, in the pre-training of LLMs, x denotes the unmasked token, while y represents the prediction target with $[CLS]$. The proposed method creates the thought chain P_t as follows:

$$\{e(h_0 : m), e(s_0 : n), e(p_0 : m), e(y)\} \tag{12}$$

where $h_{0:m}(0 < i < m)$ represents the pre-generated template with [MASK] label of length m, $s_{0:n}$ represents the n-words sentence, and $p_{0:m}$ represents consecutive pseudo-tokens that have the same length as the pre-generated template.

In NLP and LLM research, an input prompt x typically consists of an input instance and a set of potential labels Y. The language model then generates a sequence of tokens that either includes the label y as a prefix or suffix. This process can be expressed mathematically as follows:

$$y = argmax\ p(y|x, Y; \theta) \tag{13}$$

where $p(y|x, Y; \theta)$ denotes the probability of the label y given the input instance x and the set of possible labels Y. θ refers to the model parameters. The label y with the highest probability is selected using the $argmax$ operator.

The classification tasks in our study employ the use of cross-entropy loss, which can be expressed mathematically as:

$$L(\theta) = -\sum log\ p(y_i|x_i, Y; \theta) \tag{14}$$

where x_i denotes the ith input instance, y_i represents the corresponding true label, and $p(y_i|x_i, Y; \theta)$ denotes the probability of the true label y_i given the input instance x_i and the set of possible labels Y.

BERT can be applied to classification tasks by adding a classification layer on top of the pre-trained model. This layer takes the output embeddings of the final token of the input sequence as input and generates a probability distribution over the possible classes. To fine-tune BERT for a particular classification task, prompt learning can be utilized. This approach involves providing a prompt that includes both the input instance and the set of possible labels to the model. The input prompt is concatenated with the input instance and then passed through the BERT model to generate a representation for the input sequence. The representation of the final token is then used to predict the class label by passing it through the classification layer. Specifically, given an input instance x_i and a set of possible labels Y, the input prompt $\mathcal{T}$ can be defined as:

$$\mathcal{T} = [CLS]\ x_i\ [SEP]\ Y\ [SEP] \tag{15}$$

where $[CLS]$ and $[SEP]$ are special tokens used to indicate the start and end of the input sequence and the set of possible labels, respectively. We measure the difference between the predicted class

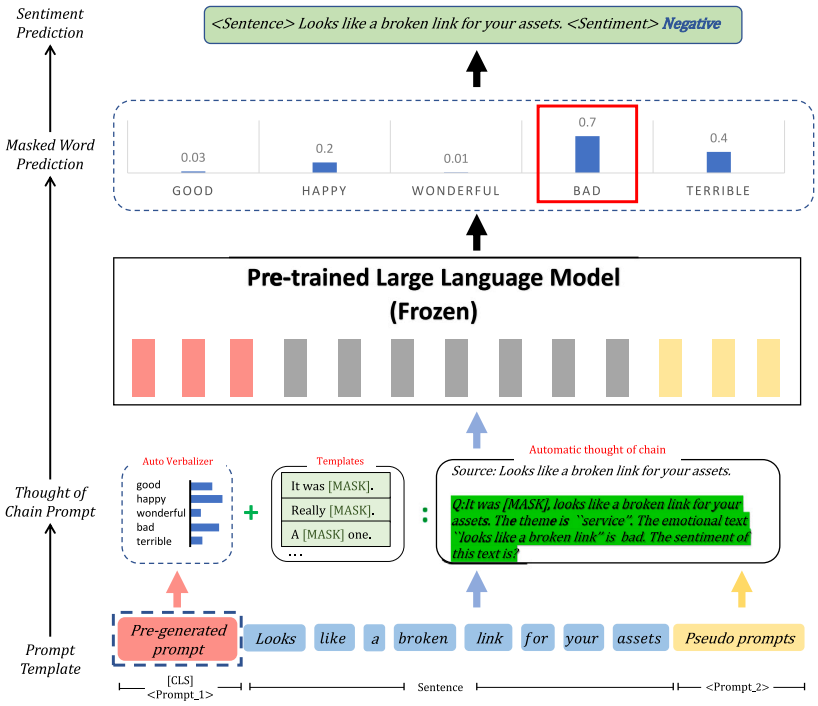


Fig. 7. Architecture of the AGCVT-Prompt for sentimental classification.

probabilities and the true class labels using the cross-entropy loss, which can be formulated as shown in Eq. (14), where y_i represents the true class label for the ith input instance x_i, $p(y_i|x_i, Y; \theta)$ represents the predicted probability distribution over the possible classes, and θ represents the model parameters. The goal is to minimize the cross-entropy loss to improve the accuracy of the classification model.

When utilizing T5 models for sentiment classification tasks, an output can be generated through the following formulation:

$$output = T5(\text{“}classify: \text{”} + \text{“}prompt\text{”} + < extra\_id\_0 > + input\_text) \quad (16)$$

To utilize T5 models for sentiment classification tasks, a specific prompt customized to the task at hand is created using the keyword “classify” at the beginning and the “prompt” variable. For example, a prompt could be “Is this a positive or negative sentence?” The “input_text” variable represents the comment sentence to be classified. The token “<extra_id_0>” serves as a distinct marker to separate the prompt from the input text, allowing the model to differentiate between the two.

4. Experiment settings

This section presents a comprehensive overview of our experimental preparations to ensure the reproducibility of our results, including datasets, parameter settings, pre-train language models, and baselines.

4.1. Datasets

Table 2 presents essential information regarding the datasets. Our experimental setup involved utilizing three datasets for binary sentiment classification, comprising one in English and two in Chinese. To evaluate the advantages of prompt learning, we restricted the dataset size. The datasets were collected from actual network platforms and made available online. Positive sentiments were labeled as positive, while negative sentiments were labeled as negative. We utilized Airline-tweet to create two different dataset sizes to comprehensively compare prompt learning and traditional fine-tuning. We procured the Chinese corpus data HoCom and WmCom from a hotel and a food delivery platform, respectively. Additionally, two different sizes of datasets from the food delivery platform are created to enable informative comparisons. SMP2020-EWECT and JPWB are six-class datasets of Chinese sentiment. The content of SMP2020-EWECT is the topic of COVID-19 on Weibo, and JPWB is the content about Japan’s nuclear sewage discharge on Weibo. The six categorical categories are Happiness, surprise, fear, sadness, anger, and disgust.

4.2. Parameter setting

Numerous parameters can impact the experiment’s results. Table 3 illustrates the primary configuration of the training components utilized in our model, along with corresponding descriptions for each configuration item. Moreover, we performed an extensive analysis of the impact of token numbers on our model’s performance through a separate hyperparameter analysis of soft prompt token numbers. This enabled us to investigate the causes underlying how varying token numbers can affect our models.

4.3. Pre-train language model

There are currently several large pre-trained models available, each capable of handling prediction tasks using various architectures. While BERT is predominantly employed for tasks like text classification and entity recognition, its efficacy in addressing intricate problems can be enhanced through the incorporation of COT prompts into the input. For instance, in sentiment analysis or text inference assignments, a series of inferential steps can be integrated into the input text. Owing to its text-to-text transformation proficiency, T5 is aptly suited for use with COT. T5 is capable of not only generating responses but also elucidating the underlying reasoning process. This attribute renders it particularly effective for tasks necessitating comprehensive explanations or extensive reasoning chains.

A common technique when using an MLM (Devlin et al., 2018) is to mask the specific part that requires prediction. Additionally, generative models have also emerged as a viable solution for prediction or classification tasks, relying on the input text to generate predictions as model outputs. Furthermore, our study involves a comparative

Table 2
Summary of experimental datasets.

Dataset	Language	Description	Train set size	Validation set size	Test set size	Total
TweetAir	English	Tweet comments about airlines	1000	400	400	1800
TweetAir-Full	English	More data of airlines	6700	1328	1328	9356
Tweet2022	English	The texts of the tweet	1500	600	600	2700
HoCom	Chinese	Comments about hotels	1000	400	400	1800
WmCom	Chinese	Comments about delivered food	1000	300	300	1600
WmCom-Full	Chinese	More data of delivered food	8998	1500	1500	11 998
SMP2020-EWECT	Chinese	Six categories about COVID-19	8606	2000	3000	13 606
JPWB	Chinese	Six categories about nuclear pollution	450	150	150	800

Table 3
Implementation framework of the proposed model.

Parameters	Value	Description
Optimizer	$AdamW$	An improved training optimization algorithm based on $Adam$.
Loss function	$CrossEntropy$	Designed to measure the difference between two probability distributions and usually applied to classification model training.
Learning rate	$1e{-}5$	The hyperparameter that determines the speed of model training gradient descent.

analysis of various LLMs. Given that ChatGPT is limited to API calls and lacks the capability for adjustment via prompt learning methods, alternative open-source LLMs, namely LLaMA-7B-Chinese and Qwen-7B, are employed. These models have been selected due to their analogous performance characteristics and serve as substitutes in our evaluation. For our study, we used the following models:

BERT-base-cased (Devlin et al., 2018) MLMs are a dominant forecasting model in NLP. Therefore, we chose the classical BERT-based-cased model for our study. This model has an outstanding performance in solving English text and serves as the backbone network for various NLP tasks. It is based on the encoder structure of Transformer (Vaswani et al., 2017). BERT-base-cased has reached 110M parameters.

BERT-base-chinese (Devlin et al., 2018) The BERT-base-chinese model, developed and trained by Google, has been widely used in Chinese NLP tasks. In our study, we utilized this model to process our Chinese datasets.

T5-base (Raffel et al., 2019) Among the various language models available, GPT-2 (Radford et al., 2019) by OpenAI is one of the most well-known models, known for its strong performance. However, the left-to-right prediction approach of GPT-2 has been surpassed by the T5 series, which was recently proposed by Google researchers. T5 utilizes a lot of prompt learning in its training, which has resulted in significant performance improvements. For our experiments on English datasets, we adopted the T5-base model with 2.2B parameters to evaluate the effectiveness of our proposed method.

MT5-small (Xue et al., 2020) MT5 is the multilingual version of T5, which was developed and released by Google. In particular, the MT5-small variant is suitable for Chinese datasets, as it has lower parameters (60M) compared to other versions.

LLaMA-7B-Chinese (Touvron et al., 2023) Meta AI achieves the best possible performance under different inference budgets by training on more tokens than typically used. The resulting model, called LLaMA, has parameters ranging from 7B to 65B, making it competitive with the best existing LLMs.

Qwen-7B (Bai et al., 2023) Qwen-7B is a 7b-parameter model of a large series of pre-trained models released by Aliyun.

Table 4
Confusion matrix.

Actual Class	Predicted class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Falcon (Penedo et al., 2023) Falcon-40B is the first open-source model with capabilities comparable to many current closed-source models, created by the Technology Innovation Institute (TII)

We deployed parts of our experiments that relied on Openpromt tools (Ding et al., 2021).

4.4. Metrics

To assess the effectiveness of our proposed model in sentimental classification, we consider the following performance metrics: Accuracy score (Acc) and F1 score (F1). We divide the experiment results into four categories according to Table 4. Accuracy, precision, recall and F1-score are defined as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ F1 &= \frac{2*Precision*Recall}{Precision+Recall} \end{aligned} \tag{17}$$

During the pre-generation template stage, the fluency of prefixes is evaluated using the bleu index. The scoring formula for the bleu index is defined as follows:

$$Bleu_n = \frac{\sum_{c\in candidates}\sum_{n-gram\in c} Count_{clip}(n-gram)}{\sum_{c'\in candidates}\sum_{n-gram'\in c'} Count_{clip}(n-gram')} \tag{18}$$

Our T5-generated templates are represented as $candidates$, while the entire training corpus is represented as $reference$. The $\sum_{c\in candidates}$ term counts all the candidates, and the $\sum_{n-gram\in c}$ term counts all n-grams in a candidate. $Count_{clip}(n-gram)$ indicates the number of n-grams in the reference. The numerator in Eq. (18) counts how many n-gram words appear in the reference for a given candidate, while the denominator represents the total number of n-grams in all candidates.

4.5. Pre-generated prompt

Our proposed approach adopts a pre-generated hard template methodology, which involves training T5-type models on the TweetAir-Full and WmCom-Full datasets. The quality of the generated templates was assessed using a synergistic score that incorporates three metrics: accuracy and F1 scores on our datasets, as well as a sentence's readability score. The readability of the templates was evaluated using BLEU (Papineni et al., 2002), which measures the fluency of a sentence. Fluent and readable sentences are known to assist the model in

Table 5
Performance scores for pre-generated prompt templates.

Template	Language	Acc	F1	Readability	Weighted Sum
It was [MASK].	English	92.27	90.83	95.00	**92.21**
In [MASK].	English	92.50	91.44	80.00	89.56
A [MASK] one.	English	88.62	86.56	90.00	88.07
Is [MASK].	English	89.74	88.30	75.00	86.22
这家很	Chinese	91.57	88.65	95.00	**91.01**
是很	Chinese	90.62	88.78	80.00	87.76
这是	Chinese	92.24	91.05	75.00	88.32
我来	Chinese	91.48	89.76	80.00	88.50

reasoning through a chain of thoughts, as pointed out by Wei et al. (2022). The high-scoring templates are displayed in Table 5. We select "it was [MASK]" and "这家很" as the hard template part of our method after ranking the three scores by weight. Although these two templates do not perform best on their own dataset, they are more readable in the context. Considering judging an item in English, "It was great. Flight attendants are gentle." is obviously much more fluent than "In great. Flight attendants are gentle". When a complete sentence "Looks like a broken link for your assets." is input, the English model will automatically generate as "It was [MASK]. Looks like a broken link for your assets. Soft Tokens". Similarly, the Chinese model generates as ""这家（外卖）很[MASK]。肉的分量太少了，速度也很慢。"

We have incorporated "it was [MASK]" and "这家很" as the hard template part of our method, based on their ranking in terms of the three scores with appropriate weighting. Although these two templates do not perform the best on their respective datasets, they are more readable and coherent in context. For example, when evaluating an item in English, "It was great. Flight attendants are gentle" is more fluent than "In great. Flight attendants are gentle". When a complete sentence, such as "Looks like a broken link for your assets", is inputted, the English model generates "It was [MASK]. Looks like a broken link for your assets. Soft Tokens" automatically. Similarly, the Chinese model generates as ""这家（外卖）很[MASK]。肉的分量太少了，速度也很慢。".

4.6. Baselines

To improve the scalability of our proposed prompt learning method for sentiment classification, we compared our AGCVT-Prompt method with recent methods. Additionally, we conducted a comparative analysis of different existing prompt configurations to evaluate our model and conduct more comprehensive research.

RecogNet-LSTM+CNN (Ramaswamy and Jayakumar, 2022) Ramaswamy et al. integrated explicit knowledge from an external database (RecogNet) with implicit information from an LSTM model and implemented a CNN with an object and position attention mechanism on the RecogNet-LSTM layer.

BERT+BiGRU (Zhang et al., 2023) GRUs are a type of RNN variant. Despite having a simpler structure, GRUs are as effective as LSTMs in handling time series. Zhang et al. incorporated BERT embeddings and employed a bi-directional GRU for classification purposes.

Bert-BiCNN (Kaur and Kaur, 2023) Two CNNs are used to extract the sentimental features of the text, and then Bert architecture is used for feature analysis and sentiment classification.

HybridBert (Talaat, 2023) Multiple Bert models are integrated into a single system and the most suitable architecture for the downstream task is selected.

Fine-tuning (FT) (Devlin et al., 2018) The traditional fine-tuning method inputs the hidden embedding of [CLS] token of the LLMs into the classification layer to make predictions. Since the classification layer is randomly initialized, fine-tuning cannot be applied to the zero-shot setting.

Pattern-Exploiting training (PET) (Schick and Schütze, 2021) The conventional prompt-tuning method, which PET adopts and most prior works, employs the class name as the sole label word for each category. To unlock the potential downstream task ability of pre-trained models, PET facilitates semi-supervised training through artificial prefix templates.

P-tuning (PT) (Devlin et al., 2018) Unlike PET's hard template, P-tuning initializes prompt templates with pseudo-tokens in an unsupervised training manner.

5. Result

5.1. Comparative experiment analysis

In this section, we outline our experimental design and results. We abstain from utilizing common training techniques such as self-training and prompt ensemble (Schick and Schütze, 2021), which are often employed in prompt learning experiments. Separate contrast experiments were conducted on the English and Chinese datasets, and the obtained results are presented in Table 6.

Our proposed approach, which is based on BERT and T5 LMs, demonstrates superior performance on two English and two Chinese datasets. These results provide strong evidence for the effectiveness of prompt learning methods in sentiment analysis tasks. Guided by the COT, our model accurately determines the sentiment polarity of complex semantic structures in both Chinese and English sentences. Liu et al. (2021) highlight the potential of LLMs in tackling downstream tasks, and our work is inspired by their findings. To explore the potential of prompt learning methods in zero-shot classification tasks, we conducted experiments on the same datasets without training the model. These zero-shot experiments with different prompt templates are presented in Table 7.

Our approach uses pre-trained models with different templates, and we only adopt semi-supervised training methods. In the zero-shot experiment, we evaluate the model without any additional training, and we observe that different prompt templates lead to different performances under the same LLM. We did not use mean and standard deviations since the results of the zero-shot experiment remain unchanged when the prompt template remains the same. Our proposed method still achieves excellent performance even in the zero-shot setting. Specifically, our approach based on BERT models achieved the best results on two English datasets and one Chinese dataset. Similarly, for T5-based models, our approach achieved the best results on all four datasets.

The results presented in Tables 6 and 7 demonstrate that prompt learning methods can achieve strong performance on sentiment classification tasks, even when provided with only limited amounts of data. Notably, some of the models were able to achieve accuracies of 60%–80% even in the absence of any training data.

Table 6
Recent competent baselines for classification. Four prompt learning techniques using BERT and T5 as pre-trained models: Fine-Tuning (FT), Pattern-Exploiting Training (PET), P-Tuning (PT), and the proposed AGCVT-Prompt method.

Approaches	Datasets and metrics							
	TweetAir		Tweet2022		HoCom		WmCom	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RecogNet-LSTM+CNN	74.32 ± 0.72	72.44 ± 0.47	75.02 ± 0.53	74.08 ± 0.22	58.82 ± 0.14	56.05 ± 0.33	62.03 ± 0.10	58.32 ± 0.26
BERT+BiGRU	77.03 ± 0.26	76.44 ± 0.37	80.65 ± 0.83	78.84 ± 0.26	75.34 ± 0.26	74.72 ± 0.39	80.50 ± 0.66	78.42 ± 0.05
BERT(FT)	78.80 ± 0.55	77.47 ± 0.35	78.12 ± 0.04	78.55 ± 0.20	80.86 ± 0.18	78.30 ± 0.45	81.67 ± 0.28	79.10 ± 0.67
BERT(PET)	94.02 ± 0.33	92.25 ± 0.46	91.88 ± 0.24	88.31 ± 0.64	85.24 ± 0.38	82.35 ± 0.45	91.25 ± 0.43	90.31 ± 0.39
BERT(PT)	81.00 ± 0.63	76.22 ± 0.65	89.78 ± 0.14	87.60 ± 0.39	73.64 ± 0.37	68.31 ± 0.67	74.50 ± 0.96	73.48 ± 0.89
BERT(Ours)	**94.22 ± 0.18**	**93.18 ± 0.22**	**92.47 ± 0.58**	**92.38 ± 0.43**	**85.28 ± 0.45**	**83.30 ± 0.68**	**92.50 ± 0.50**	**91.32 ± 0.14**
T5(FT)	87.58 ± 0.62	85.40 ± 0.74	79.99 ± 0.67	78.25 ± 0.14	81.95 ± 0.77	80.30 ± 0.59	80.38 ± 0.06	79.60 ± 0.92
T5(PET)	90.82 ± 0.35	88.57 ± 0.53	92.13 ± 0.45	90.29 ± 0.28	92.55 ± 0.16	90.95 ± 0.31	94.37 ± 0.38	93.14 ± 0.17
T5(PT)	93.15 ± 0.60	92.30 ± 0.49	92.46 ± 0.18	90.57 ± 0.25	92.24 ± 0.78	91.27 ± 0.02	92.62 ± 0.46	90.92 ± 0.59
T5(Ours)	**94.95 ± 0.29**	**93.17 ± 0.34**	**92.77 ± 0.21**	**90.62 ± 0.44**	**93.22 ± 0.14**	**91.50 ± 0.36**	**94.80 ± 0.47**	**94.55 ± 0.41**

Table 7
Zero-shot evaluation of various prompt templates.

Approaches	Datasets and metrics							
	TweetAir		Tweet2022		HoCom		WmCom	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT(PET)	0.53	0.45	0.42	0.36	**0.75**	**0.68**	0.64	0.43
BERT(PT)	0.41	0.40	0.41	0.35	0.68	0.55	0.34	0.25
BERT(Ours)	**0.78**	**0.66**	**0.51**	**0.44**	0.74	0.67	**0.67**	**0.48**
T5(PET)	0.50	0.46	0.43	0.38	0.60	0.58	0.66	0.62
T5(PT)	0.56	0.54	0.42	0.36	0.50	0.47	0.55	0.50
T5(Ours)	**0.64**	**0.60**	**0.54**	**0.46**	**0.62**	**0.59**	**0.69**	**0.67**

Table 8
Six thought chains of AGCVT-Prompt model architectures.

Index	Abbreviation	Sequence
I	P-T-S	pre-generated template/text/soft prompt tokens
II	P-S-T	pre-generated template/soft prompt tokens/text
III	T-P-S	text/pre-generated template/soft prompt tokens
IV	T-S-P	text/soft prompt tokens/pre-generated template
V	S-P-T	soft prompt tokens/pre-generated template/text
VI	S-T-P	soft prompt tokens/text/pre-generated template

5.2. Macroscopic arrangement analysis

The p-tuning method appears to perform poorly in Chinese sentiment classification when compared to the PET method and our proposed AGCVT-Prompt method, as we have observed. Our proposed chain of thought method is also based on a continuous or discrete prompt. We have further noticed that when the effect of p-tuning is much better than that of PET, AGCVT also outperforms PET. However, when p-tuning performs poorly, the improvement of AGCVT-Prompt over PET is also insignificant. Our COT can be broken down into three parts, which are a pre-generated template, an input sentence, and pseudo-tokens used to represent a soft template, before embedding the LLMs.

The relative positions of the three parts of AGCVT-Prompt, namely the pre-generated hard template, the combined generated content and comment text, and the soft prompt tokens, were found to have an impact on the performance of prompt learning. To provide a comprehensive understanding of the results, these three parts were permuted into 6 cases, which are presented in Table 8.

The performance of the six combinations on four datasets was tested using six different thought chains. Fig. 13 displays the results. The chart demonstrates that the combination of the P-T-S chain results in the best performance. This permutation inputs a fixed prefix before the text, providing LLMs with enough clues for downstream tasks. The model optimizes the soft prompt tokens on the training set, and these tokens, in turn, supplement the pre-generated templates. This chain of thought enables models to deduce backwards after forward reasoning, thereby improving accuracy.

The combination of the T-P-S chain performs similarly to the PET method, with only minor improvements. This chain is equivalent to the quadratic reasoning of the PET method. AGCVT-Prompt performs about the same as the p-tuning method when applying sequences T-S-P and S-T-P. In these two combinations, soft templates play a dominant role in text semantics. The T-S-P and S-T-P structures allow LLMs to focus mainly on pseudo tokens while training the parameters for prompts. LLMs cannot reason in the chain of thought under these two combinations. The weak results of the P-S-T and S-P-T combinations indicate that placing the text at the end is not conducive to LLMs classification.

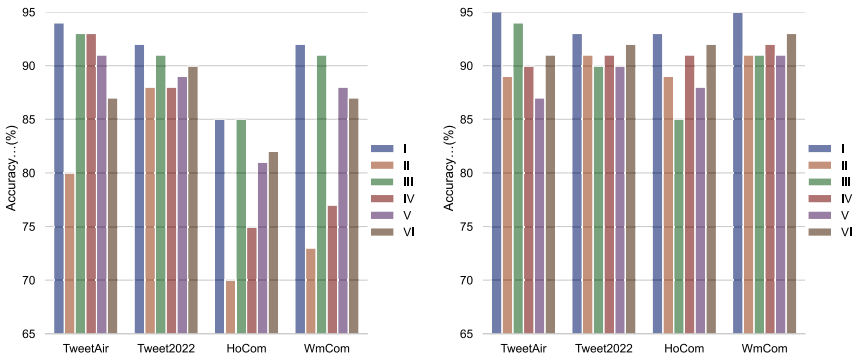
5.3. Six-categories classification analysis

Expanding from the primary binary sentiment classification, we developed a six-category sentiment classification framework. This involved using two novel datasets annotated to reflect six primary sentiments: anger, disgust, fear, happiness, sadness, and surprise. These sentiments, characterized by their distinct attributes, allow for diverse degrees of expression. In this context, sentiments are categorized discretely rather than reflecting individual sentimental states. Given the inadequacy of binary classifiers for this multi-category task, all models required retraining. We employed the Bert-base-Chinese model, retrofitting it for the six-category task by retraining the classification head through fine-tuning and prompt-based learning techniques (see Fig. 8).

Further, we assessed the performance of two recent BERT methodologies, Bert-BiCNN and HybridBert. Additionally, we trained the LLaMA, Qwen, and Falcon models using p-tuning and Lora methods, conducting experiments on the datasets above. The sentiment flow classification in QA-based large models involved specifying output categories via prompts. Concurrent few-shot learning experiments with a 12-sample set were conducted across different methods. As indicated in Table 9, there was a notable improvement in performance across models when applied to the more extensive SMP2020 dataset, with the 7B model achieving state-of-the-art results. However, the 40B model exhibited underfitting, likely due to data limitations. In the JPWB dataset, our proposed model surpassed the performance of all models with fewer than 1B parameters and outperformed the larger Falcon model. Even under few-shot conditions, our approach maintained superior performance among sub-1B models, with only marginal differences compared to the 7B model. Consequently, AGCVT-Prompt demonstrated its effectiveness among models with millions of parameters, showing only slight disparities when compared to fine-tuned LLMs.

5.4. Few-shot and zero-shot analysis

The study involves conducting few-shot learning experiments on two datasets, namely TweetAir-Full and WmCom-Full, while considering the amount of data as a controllable variable. A comparison is made between the traditional fine-tuning method and several prompt learning methods. Fig. 9 presents the results of the experiments.



(a) BERT-based models (b) T5-based models

Fig. 8. Accuracy of different prompt template combinations on diverse datasets.

Table 9
Results of sentiment six classification and few-shot experiments.

Approaches	Parameters	SMP2020		JPWB		SMP2020(12-shot)		JPWB(12-shot)	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bert-p-tuning	0.07M	0.65	0.58	0.55	0.48	0.52	0.44	0.56	0.52
T5-p-tuning	0.07M	0.83	0.77	0.69	0.62	0.55	0.48	0.60	0.56
Bert-AGCVT-Prompt(**Ours**)	0.07M	0.88	0.85	0.79	0.66	0.76	0.74	0.68	0.62
T5-AGCVT-Prompt(**Ours**)	0.07M	0.90	0.87	0.72	0.59	0.79	0.75	0.69	0.66
Bert-fintuning	300M	0.89	0.82	0.42	0.36	0.39	0.37	0.38	0.29
Bert-BiCNN	300M	0.85	0.83	0.51	0.44	0.36	0.32	0.40	0.35
HybridBert	500M	0.90	0.84	0.48	0.46	0.33	0.26	0.28	0.23
LlaMA-p-tuning	7M	**0.96**	**0.94**	0.83	0.82	0.82	0.77	0.71	0.66
Llama-Lora	7B	0.91	0.88	0.76	0.74	0.78	0.75	0.66	0.58
Qwen-p-tuning	7M	0.93	0.90	**0.84**	**0.81**	**0.85**	**0.82**	**0.79**	**0.76**
Qwen-Lora	7B	0.89	0.82	0.79	0.75	0.80	0.77	0.75	0.73
Falcon-p-tuning	40M	0.84	0.80	0.75	0.72	0.72	0.68	0.66	0.64
Falcon-Lora	40B	0.82	0.77	0.68	0.59	0.64	0.57	0.59	0.56

In the absence of data, prompt learning methods show excellent zero-shot performance, as demonstrated in Table 7. With only 0.1% of data used for few-shot testing, prompt learning methods present significant advantages, and our proposed AGCVT method outperforms in both zero-shot and few-shot learning. The experimental results demonstrate that the COT method can greatly enhance LLMs' ability in few-shot and zero-shot learning. As the amount of data increases, the traditional fine-tuning method's accuracy begins to improve, gradually approaching the performance of these prompt learning methods. When disregarding the training cost and data quantity, traditional fine-tuning models are capable of achieving better results, while prompt learning methods are more appropriate for scenarios with limited data.

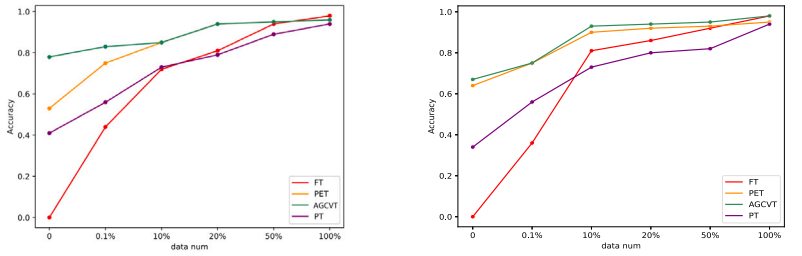
An analysis of the experimental results using 0.1% of the data, as presented in Fig. 9, demonstrates that AGCVT-Prompt significantly improves performance in few-shot prompt scenarios by automating the generation of COT. AGCVT-Prompt enhances the model's ability to comprehend and address complex problems by integrating intermediate reasoning steps such as topic and sentiment prompts. This methodology substantially improves the model's understanding in data-scarce environments, particularly in few-shot learning. By guiding the model through logical reasoning sequences, AGCVT-Prompt enhances accuracy in few-shot configurations, enabling models to make more effective inferences and generalizations even with limited data. Moreover, AGCVT-Prompt contributes to the explainability of the model, as the delineated reasoning steps provide clarity and transparency in the model's decision-making processes.

5.5. Complexity and overhead analysis

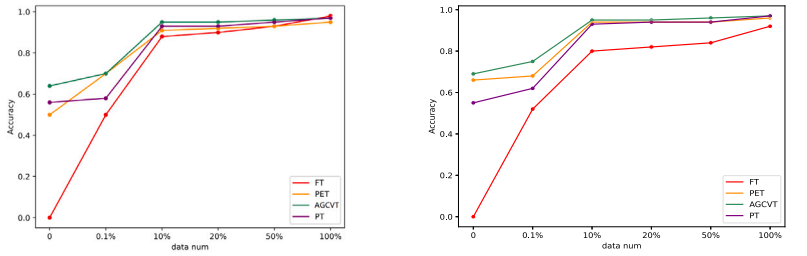
The computational time costs of various prompt learning methodologies, including FT, PET, PT, AGCVT-Prompt, and Lora for LLMs are depicted in Fig. 10. During fine-tuning, the entire parameter set of the BERT model denoted as N and approximately 110M for BERT-base, is updated. The computational complexity of BERT is primarily contingent upon its depth, hidden size, and the number of attention heads, with the complexity per layer approximating $O(n^2 \cdot d)$, where n represents the sequence length and d, the hidden size (set at 768 in our experiments). In fine-tuning, every parameter is updated at each training step, leading to an update cost of $O(N)$.

For PET and PT methods, only a fractional subset of parameters undergo training. The total parameters in the prefix are denoted as P, with $P \ll N$. Adding p prefix tokens per layer, mirroring the dimensionality of BERT's hidden size (e.g., 768 for BERT-base), results in P being approximately $12\times768\times p$. Training predominantly updates prefix parameters, thereby concentrating computation on handling prefixes. However, the overall complexity remains inferior to full fine-tuning since the BERT body remains static. The update overhead focusing on prefix parameters is $O(P)$, markedly less than that of full fine-tuning's $O(N)$. PT incurs slightly higher costs than PET due to the training of additional soft tokens.

AGCVT-Prompt comprises four steps: BERT feature extraction with complexity $O(n^2 \cdot d)$, PCA dimension reduction with complexity $O(m^2 \cdot d)$ where m is the number of data points, HDBSCAN clustering with $O(n \log n)$, and training the reasoning chain LM, akin to the PT method, with $O(P)$, substantially lower than Fine-Tuning's $O(N)$. Both PT and Lora methods for LLMs maintain the original parameters and fine-tune a minimal amount of new parameters. However, Lora does not increase inference time, rendering it marginally faster than PT. The overall training complexity is influenced by sequence length (n), hidden size (d), and depth (l), with a magnitude of $O(l \cdot n \cdot d^2 + l \cdot d \cdot n^2)$. Synthesizing experimental findings from this section, AGCVT-Prompt demonstrates an average complexity, rapid response, minimal parameter utilization, and optimal efficacy among LMs.



(a) BERT applying for TweetAir-Full. (b) BERT applying for WmCom-Full.



(c) T5 applying for TweetAir-Full. (d) T5 applying for WmCom-Full.

Fig. 9. The BERT-based and t5-based models are used for experiments on two large-scale datasets.

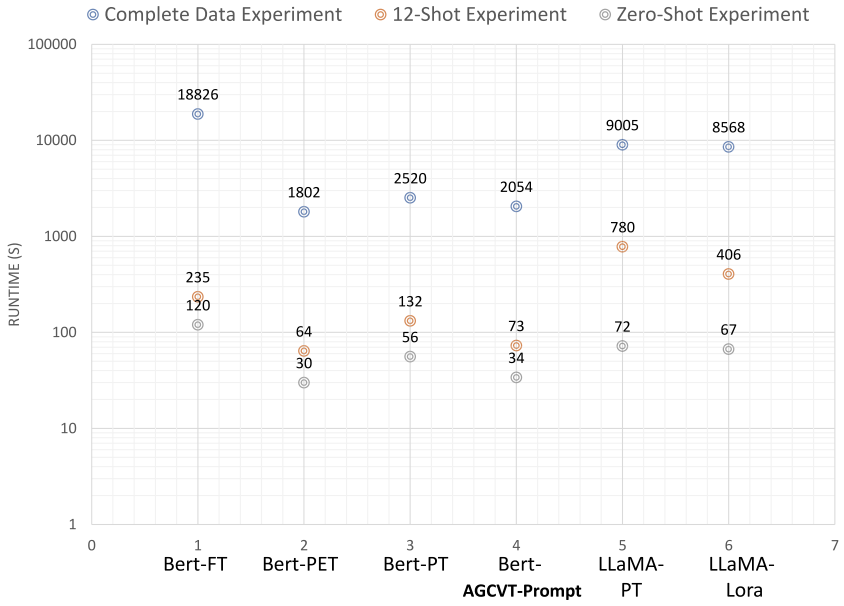
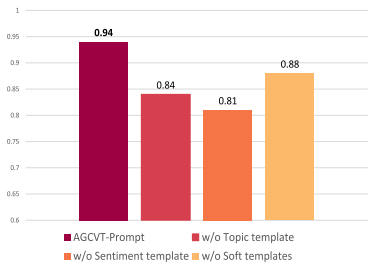


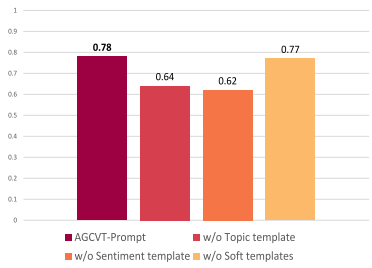
Fig. 10. Time cost of Bert and LLaMA using different training methods.

5.6. Abaltion experiment analysis

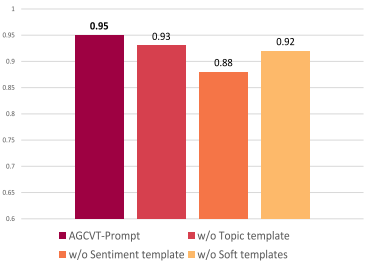
Tables 8 and Fig. 13 present investigations into the impact of varying configurations of COT components on LMs in classification tasks. To delve deeper into the contributions of individual AGCVT-Prompt modules to overall performance, ablation studies were conducted using the TweetAir dataset, which produced optimal results. The ablation experiments evaluated the AGCVT-Prompt model under various conditions: without topic prompt templates, without sentiment prompt templates, and soft templates, also within a zero-shot setting. According to the accuracy scores depicted in Fig. 11, excluding topic prompt templates resulted in performance declines of 11.9% and 21.5% in BERT for classification and zero-shot metrics, respectively, and reductions of 2.2% and 6.7% in T5. This underscores the significance of topic prompt templates. The removal of sentiment prompt templates led to decreases of 15.5% and 25.8% in BERT metrics for classification and zero-shot and 8.0% and 10.3% in T5, highlighting the critical role of sentiment prompts. The absence of soft templates accounted for 6.8% and 1.3% drops in BERT for classification and zero-shot and 3.2% and 0.01% in T5, confirming their utility in supervised classification. These experiments demonstrate that both topic and sentiment prompt templates are instrumental in leveraging the inherent linguistic capabilities of LMs for downstream tasks, with sentiment templates playing a more pivotal role. The results also indicate that generating COT prompts is more effective than generating individual topic or sentiment prompt templates. Given the lack of backward propagation in zero-shot learning, pseudo tokens do not update, offering any benefit for task completion and may even negatively impact model performance.



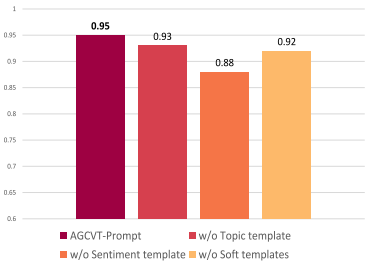
(a) Ablation experiments using Bert.



(b) Ablation experiments using Bert for zero-shot test.



(c) Ablation experiments using T5.



(d) Ablation experiments using T5 for zero-shot test.

Fig. 11. Ablation experiments for AGCVT-Prompt using Bert model and T5 model.

5.7. Prompt token analysis

Ablation studies have elucidated that the length of AGCVT-Prompt and the token length of prompt terms significantly influence the efficacy of prompt-based paradigms. Therefore, we conducted comparative experiments involving LMs employing p-tuning and AGCVT-Prompt and leading multi-turn conversational LLMs utilizing COT as outlined in Table 9. The experiments varied the number of tokens from p-tuning, AGCVT-Prompt, and COT, with this variation quantified by the aggregate of input prompt and model output tokens. Since LMs like LLaMA accommodate up to 2000 tokens, our control range was set between 10 to 1000 tokens to avert truncated outputs. For Bert, which has a limitation of 512 tokens, a sliding window technique was applied for lengths surpassing this threshold. As illustrated in Fig. 12, where darker hues represent higher accuracy scores, Bert and T5 displayed peak performance at 200 tokens, with a noticeable decline in efficacy beyond this point. LLaMA and Qwen achieved optimal performance at 400 tokens, whereas Falcon exhibited maximal effectiveness at 800 tokens. These findings substantiate that longer prompts do not invariably translate to augmented model capabilities. Particularly for LMs, restricting the token count to approximately 200 can enhance operational efficiency. Larger model sizes, conversely, benefit from an increased token allowance, which aids in learning and executing downstream tasks more effectively.

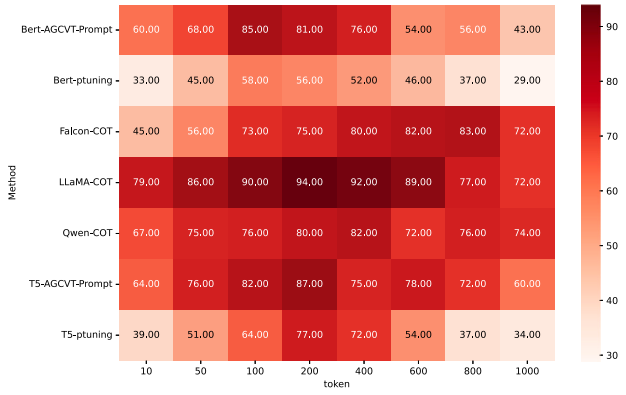


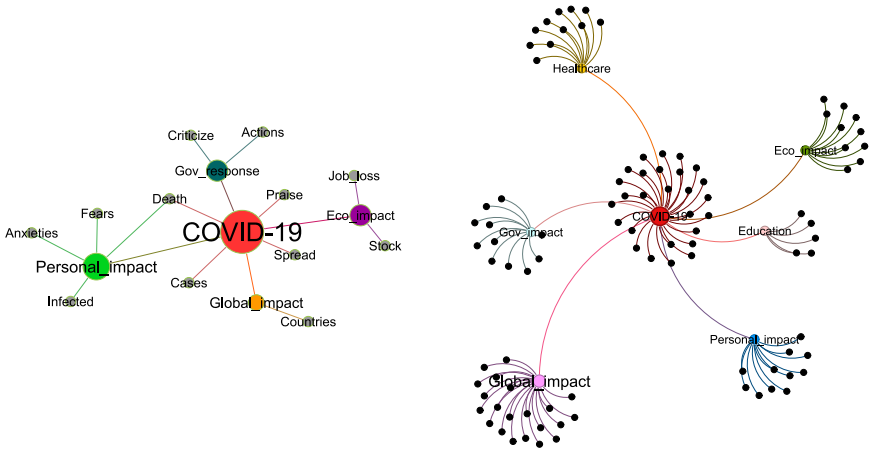
Fig. 12. F1 value heatmaps for different models with different numbers of tokens.

5.8. COVID-19 topics and sentiment analysis

An unsupervised strategy was utilized to train the pre-generated template method without using labeled data. Moreover, a generative approach was employed to enable the T5 model to cluster the data, rather than artificially providing a verbalizer. The clusters generated as a result were used as the new verbalizers, which were then integrated throughout the entire model. To generate a keyword label for each comment text, we utilized a T5 model with prompts from pre-generated templates after inputting all the sentences. These labels proved beneficial in aiding LLMs to cluster all the comments.

We collected 100 and 1000 comments related to Corona Virus Disease 2019 (COVID-19) from Tweet2022 and used them as input data. In alignment with the methodologies delineated in Eqs. (1)–(7) of Section 3.2, the verbalizer is autonomously synthesized through an initial clustering phase utilizing the HDBSCAN method as (6), followed by a secondary computation process within the T5 model. This procedure entailed extracting data from both the HDBSCAN layer and the output layer of the T5 model. The experimental results were then visualized and presented in Fig. 13.

Despite the small amount of data, our model was able to successfully identify the central topic as the COVID-19 pandemic. By incorporating COT, the model was able to further classify the topics. The resulting topics were ranked in descending order of frequency as personal impact, government impact, economic impact, and global impact. The personal influence depicted in these 100 sentences primarily pertains to individual infections and sentiments toward the pandemic. By analyzing the clusters that are most closely connected with the government, such as “criticize” and “actions”, it can be deduced that comments related to the government mainly consist of criticisms regarding their



(a) Clustering results for 100 product reviews. (b) Clustering results for 1000 product reviews.

Fig. 13. Review clustering results using the AGCVT-Prompt model.

mishandling of the pandemic and calls for the government to take action in response to the outbreak. Internet users discussing the impact of the pandemic on the economy primarily focus on the stock market downturn and unemployment.

Although the names of multiple countries appear in these reviews, the model is limited by the small amount of textual data, which prevents a more nuanced breakdown. When 1000 comments were used for analysis, the clustering effect was significantly better and more categories were divided. After expanding the comments data, the proposed model improved the clustering results. Comment topics were divided into seven categories: COVID-19 pandemic, global impact, healthcare, personal impact, government impact, economic impact, and educational impact. We used these seven classifications as verbalizers for the review topic classification task. The keywords of each review were generated by the T5 model with our prompts.

In this study, we utilized the thought chain prompts method to conduct ten rounds of classification and calculated the average as the final result. For each category, we extracted the top five keywords and used the verbalizer with the highest probability to determine the category of a sentence. It should be noted that a single sentence may have multiple themes. For example, the sentence "In the past week, there have been one million new infections globally" is tagged as "cases", but it is closely related to "pandemic" and "global impact". The classification results are visually represented in Fig. 14. This figure represents the probability that each keyword belongs to the corresponding category. The color intensity of each keyword indicates this probability. It is evident that in the pandemic topic, the most frequently mentioned keywords by netizens in comments are quarantine, public health, cases, deaths, and impact. The health topic, on the other hand, includes keywords that people focused on, such as public health, hospitalization, medications, patients, and healthcare. In the economic topic, funding, industries, stock markets, unemployment, and transportation are among the most discussed topics. Individuals' personal lives also come up in the comments, including pregnancy and childbirth during the pandemic, negative sentiments caused by the pandemic, and unemployment. When it comes to the government, the most discussed topics are the government's response, the behavior of leaders, policies, and guidelines. Additionally, some peace-loving netizens urge the US government to lift sanctions on Iran during the pandemic. The issue of Iran is also the most widely discussed global issue.

Furthermore, in the global topic, netizens discuss the cases and impacts of pandemics worldwide, particularly outbreaks. Education-related topics mainly pertain to schools. By applying our proposed chain of thought prompt and generative models, LLMs can accomplish unsupervised classification tasks and analyze the text's themes and content more deeply.

5.9. Ethics and privacy challenges

Regarding sentiment classification in healthcare, using BERT or T5 models with the AGCVT-Prompt method has advantages and considerations to remember. However, it is essential to be mindful of data protection and the potential implications of automated sentiment analysis:

Data Privacy and Sensitivity: Healthcare data is susceptible, and handling it with utmost care is crucial. Violating privacy laws can result in legal consequences and a loss of public trust. Therefore, all data processing must strictly adhere to data protection laws, and privacy protection protocols must be in place. Before model training, any potentially sensitive comments should be anonymized to maintain patient confidentiality.

Bias and Fairness: There is a risk that pre-trained models may internalize and magnify existing biases in the training data. This could lead to biased outcomes, particularly regarding gender, race, and age. We must carefully examine the training dataset to identify and rectify any inherent biases to prevent this.

Misunderstanding and Misuse: Automated sentiment analysis in healthcare may misinterpret subtle sentimental nuances in highly personalized contexts, such as patient care. Misreading patient sentiments could result in inadequate or inappropriate medical responses. It is advisable to integrate human assessment to prevent this, allowing for human-in-the-loop evaluation and explanation of automated analytical outcomes.

By implementing these strategies and safeguards, the deployment of AGCVT-Prompt in healthcare sentiment analysis can maximize its benefits while substantially reducing potential risks and negative impacts.

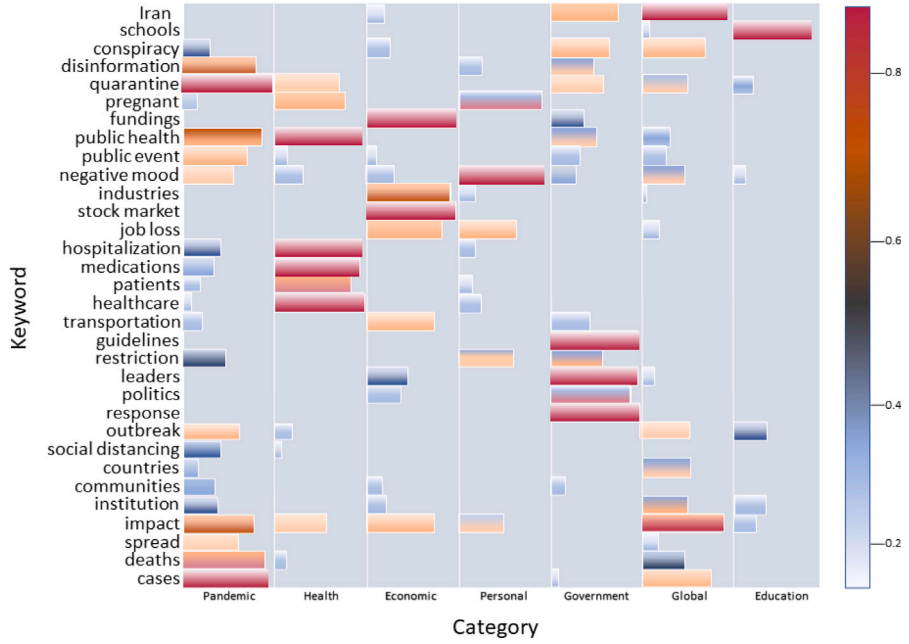


Fig. 14. Semantic keywords extracted and clustered into seven categories.

6. Conclusion and future work

This paper introduces a new algorithm for sentiment classification called AGCVT-Prompt. The algorithm only requires a small number of parameters to be trained and utilizes large language models (LLMs) to leverage their potential knowledge and guide their reasoning through a chain of thought (COT) in order to derive accurate answers. Through the use of this model to process network comments, various prompts can be utilized to classify text keywords by topic and count word frequency.

Prior research mainly employed continuous or discrete prompt templates to enable LLMs to solve downstream tasks. The quality of the template influences LLM's performance, with human-designed templates generally achieving better results than continuous pseudo-tokens with high readability. However, creating high-quality prompt templates is a challenging task. The advent of COT has bolstered LLM's reasoning ability, yet there are only a few COT prompt learning methods for sentiment classification. Our proposed method generates templates based on the T5 model to automatically generate high-quality templates by validating set indicators.

AGCVT-Prompt automatically constructs COT prompts by generating templates, text, and soft prompt tokens in advance to assist LMs in classification tasks. Prior prompt learning research did not optimize verbalizers. Therefore, we introduce a technique to generate verbalizers through T5 model prompts, which automatically creates verbalizers based on the content of the dataset. In future research, we aim to utilize the proposed method to solve additional downstream tasks and improve the soft templates.

This study compares traditional prompt learning methods with advanced sentiment classification methods using BERT and T5 models on two Chinese datasets and two English datasets, achieving superior performance on sentiment binary classification tasks. Further investigations into the more nuanced six-category sentiment classification on the Chinese datasets revealed AGCVT-Prompt's remarkable efficacy, even with significantly smaller model sizes than leading generative LLMs. Moreover, AGCVT-Prompt has demonstrated excellent zero-shot and few-shot learning capabilities. The study examined the influence of COT configurations and token lengths on model performance, aiming to delineate performance patterns of AGCVT-Prompt relative to alternative approaches. Ablation studies confirmed the effectiveness of each component within AGCVT-Prompt and highlighted the superiority of automated COT prompt generation over general prompt template construction. Solutions to ethical challenges are also discussed. Finally, using AGCVT-Prompt to extract text keywords can help researchers gain a detailed understanding of the content themes. In the future, the research aims to develop and integrate plug-and-play automated COT modules for open-source LLMs.

CRediT authorship contribution statement

Xu Gu: Conceptualization, Data curation, Formal analysis, Investigation, Software, Visualization, Writing – original draft. **Xiaoliang Chen:** Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **Peng Lu:** Validation, Writing – review & editing. **Zonggen Li:** Data curation, Resources. **Yajun Du:** Funding acquisition. **Xianyong Li:** Validation.

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

Data will be made available on request.

Acknowledgments

This work is supported by the Science and Technology Program of Sichuan Province (Grant no. 2023YFS0424) and the National Natural Science Foundation (Grant nos. 61902324, 11426179, and 61872298). X. Gu, Conceptualization; X. Gu and Z. G. Li, Data curation; X. Gu, Formal analysis; X. L. Chen and Y. J. Du, Funding acquisition; X. Gu, Investigation; X. L. Chen, Methodology; X. L. Chen, Project administration; Z. G. Li, Resources; X. Gu, Software; X. L. Chen, Supervision; P. Lu and X. Y. Li, Validation; X. Gu, Visualization; X. Gu/Writing-original draft; X. L. Chen, P. Lu/Writing-review and editing.

References

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T., 2023. Qwen technical report. arXiv preprint arXiv:2309.16609. URL https://arxiv.org/abs/2309.16609.

Brown, T.B., Mann, B., Ryder, N., et al., 2020. Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual.

Cho, K., Van, M.B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. http://dx.doi.org/10.48550/arXiv.1406.1078, arXiv preprint arXiv:1406.1078.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. http://dx.doi.org/10.48550/arXiv.1810.04805, CoRR abs/1810.04805.

Diao, S., Wang, P., Lin, Y., Zhang, T., 2023. Active prompting with chain-of-thought for large language models. arXiv:2302.12246.

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H., Sun, M., 2021. OpenPrompt: An open-source framework for prompt-learning. arXiv preprint arXiv:2111.01998.

Eldridge, J., Belkin, M., Wang, Y., 2015. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In: Proceedings of the 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015. In: JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, pp. 588–606, URL http://proceedings.mlr.press/v40/Eldridge15.html.

Gao, T., Fisch, A., Chen, D., 2021. Making pre-trained language models better few-shot learners. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, pp. 3816–3830. http://dx.doi.org/10.18653/v1/2021.acl-long.295.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2015. LSTM: A search space odyssey. http://dx.doi.org/10.1109/TNNLS.2016.2582924, CoRR abs/1503.04069.

Ho, N., Schmid, L., Yun, S., 2022. Large language models are reasoning teachers. arXiv:2212.10071.

Hu, S., Ding, N., Wang, H., Liu, Z., Wang, J., Li, J., Wu, W., Sun, M., 2022. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, pp. 2225–2240. http://dx.doi.org/10.18653/v1/2022.acl-long.158.

Kaur, K., Kaur, P., 2023. Improving BERT model for requirements classification by bidirectional LSTM-CNN deep model. Comput. Electr. Eng. 108, 108699. http://dx.doi.org/10.1016/J.COMPELECENG.2023.108699.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2019. ALBERT: a lite BERT for self-supervised learning of language representations. http://dx.doi.org/10.48550/arXiv.1909.11942, CoRR abs/1909.11942.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. 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, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, pp. 7871–7880. http://dx.doi.org/10.18653/v1/2020.acl-main.703.

Li, C., Gao, F., Bu, J., Xu, L., Chen, X., Gu, Y., Shao, Z., Zheng, Q., Zhang, N., Wang, Y., Yu, Z., 2021. SentiPrompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. CoRR abs/2109.08306. URL https://arxiv.org/abs/2109.08306.

Li, X.L., Liang, P., 2021. Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, pp. 4582–4597. http://dx.doi.org/10.18653/v1/2021.acl-long.353.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized BERT pretraining approach. http://dx.doi.org/10.48550/arXiv.1907.11692, CoRR abs/1907.11692.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J., 2021. GPT understands, too. CoRR abs/2103.10385. arXiv:2103.10385.

OpenAI, 2023. GPT-4 technical report. arXiv:2303.08774.

Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, pp. 311–318. http://dx.doi.org/10.3115/1073083.1073135.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J., 2023. The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 URL https://arxiv.org/abs/2306.01116.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training. OpenAI Blog 1 (8), 9.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P., 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints arXiv:1910.10683.

Ramaswamy, S., Jayakumar, C., 2022. RecogNet-lstm+cnn: a hybrid network with attention mechanism for aspect categorization and sentiment classification. J. Intell. Inf. Syst. 58 (2), 379–404. http://dx.doi.org/10.1007/s10844-021-00692-3.

Schick, T., Schütze, H., 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp. 255–269. http://dx.doi.org/10.18653/v1/2021.eacl-main.20.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., Zhou, D., 2023. Large language models can be easily distracted by irrelevant context. arXiv:2302.00093.

Talaat, A.S., 2023. Sentiment analysis classification system using hybrid BERT models. J. Big Data 10 (1), 110. http://dx.doi.org/10.1186/S40537-023-00781-W.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G., 2023. Llama: Open and efficient foundation language models. URL https://doi.org/10.48550/arXiv.2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. http://dx.doi.org/10.48550/arXiv.1706.03762, CoRR abs/1706.03762.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q., Zhou, D., 2022. Chain of thought prompting elicits reasoning in large language models. CoRR abs/2201.11903. URL https://arxiv.org/abs/2201.11903.

Xiang, W., Wang, Z., Dai, L., Wang, B., 2022. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022. International Committee on Computational Linguistics, pp. 902–911, URL https://aclanthology.org/2022.coling-1.75.

Xu, Z., Wang, C., Qiu, M., Luo, F., Xu, R., Huang, S., Huang, J., 2023. Making pre-trained language models end-to-end few-shot learners with contrastive prompt tuning. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023. ACM, pp. 438–446. http://dx.doi.org/10.1145/3539597.3570398.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C., 2020. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv e-prints arXiv:2010.11934.

Zhang, X., Wu, Z., Liu, K., Zhao, Z., Wang, J., Wu, C., 2023. Text sentiment classification based on BERT embedding and sliced multi-head self-attention bi-GRU. Sensors 23 (3), 1481. http://dx.doi.org/10.3390/s23031481.

Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J., 2022. Large language models are human-level prompt engineers. http://dx.doi.org/10.48550/arXiv.2211.01910, CoRR abs/2211.01910.