Automated Comment Generation Based on the Large Language Model

Kaiwei Cai, Junsheng Zhou(✉), Li Kong, Dandan Liang, and Xianzhuo Li

Nanjing Normal University, Nanjing 210023, JS, China
zhoujs@njnu.edu.cn, {kongli,xli}@nnu.edu.cn

Abstract. Automated essay comment generation is important for writing in education, as it can reduce the burden on teachers while providing students with rapid feedback to improve their writing. When writing, students tend to take writing thought ideas in several excellent essays as references, which inspire them to be more creative in writing. For teachers, they can refer to them to make suggestions on how to improve the quality of essays. Inspired by this behaviour of writing psychology, we guide the Large Language Model (LLM) to also imitate this behaviour. In this paper, we apply the essay comment generation task in application, which aims to generate comments on how to amplify the writing thought ideas for the given English expository essay. To tackle this task, we propose the two-stage comment generation framework, in which we first search for some cases of the chain of writing thought ideas, and then use them as evidence to guide the LLM to learn from these references and generate more concrete comments. What's more, we manually collect some English expository essays to build the knowledge base and some essay-comment pairs (The source code is available at https://github.com/CarryCKW/EssayComGen). Extensive experiments show that our method outperforms strong baselines significantly.

Keywords: Automated essay comment generation · Large language model · Writing psychology · Knowledge base

1 Introduction

Automated essay evaluation is to evaluate the writing quality of the essay with the help of computer technologies. As a useful educational application of natural language processing (NLP), automated essay comment generation has significant social value for education. In particular, the generation of reasonable comments can reduce teachers' workload and improve teaching efficiency. What's more, students can also receive timely feedback and improve their writing skills.

Recent research [2,13,15] has demonstrated the strong performance of the Large Language Model (LLM) in NLP tasks that generate answers to user's questions without any additional fine-tuning operations, also known as

W. Hong and G. Kanaparan (Eds.): ICCSE 2023, CCIS 2023, pp. 283–294, 2024.
https://doi.org/10.1007/978-981-97-0730-0_25

LLMprompting [8]. We can utilize the strong performance of LLM in text generation and local reasoning to further advance the task of automated essay comment generation. For example, in the actual application scenario, the user will fill in the content of the essay in the following template as input to the LLM: "Please now, as an English writing tutor, make some suggestions on how to improve the quality of the essay from the perspectives of amplifying writing thought ideas. The essay: the specific content of the essay", and then the LLM will return some texts of suggestions.

However, on the one hand, a better performance of LLM always requires a higher expensive deployment cost, depending on the model parameter scales, which is not conducive to the practical generalization of LLM. On the other hand, the output of the amiable parameter of LLM might be outdated, incomplete and incorrect, and they generate factually wrong answers, called *hallucination* [9,12].

In this work, we first make a study of automated comment generation for English expository essays, which requires commenting in natural language on how to improve a given essay, because expository essay writing ideas are more structured and diverse. The challenges of this work mainly lie in the following two folds: (1) Overcoming the problem of generating incorrect answers due to the weak performance of LLM with fewer parameters, and fine-tuning the model to update the knowledge is often expensive [5], how to introduce corrective steps to inject new knowledge to improve the accuracy of answers and the reasoning performance of the model. (2) How to guide LLMs to make specific suggestions for amplifying the writing thought ideas in given essay. Overcoming these challenges can lead to the generation of better quality comments, making this research and application more valuable in practice.

When writing, students tend to learn and associate writing thought ideas from several excellent essays as a way of inspiring themselves to produce a more polished and fulfilling piece of writing. This exercise of creativity by imitating ideas of good essays can be seen as a form of inspiration and stimulation for writing, helping them to improve their writing skills and expressive abilities, and is often referred to as "mimicry" in the psychology of writing [3,4].

To tackle the above problems, inspired by the "mimicry", we propose a two-stage comment generation framework based on the search and prompting. This framework consists of two components. Specifically, the first component mimics the behaviour that students will be inspired by the writing thought ideas of high-quality essays when they are writing and expand their own writing content by associating several cases of excellent writing ideas. Here are the more granular steps. In searching, we first parse out some of the ideas of the input essay by LLM, and the obtain multiple writing cases as references by utilizing topic-aware retrieval algorithms in the back-end knowledge base, where the cases are great writing thought ideas under the relevant topics for the given essay. Furthermore, the second component is to integrate the writing cases and the input essay, as the input to LLM and guide it to learn new writing cases and generate more informative and diverse comments, which can be implemented by the well-designed prompting templates.

On the one hand, using the retrieval algorithm to get the writing cases from the knowledge base can better guide the model to learn writing habits in a human way of thinking, and can also update the content of the knowledge base in real time. On the other hand, the integration of high-quality writing thought cases with background knowledge into user questions can be more targeted to students' writing improvement, generating higher quality and more relevant improvement comments.

To support the evaluation of our task, we additionally construct a knowledge base of writing cases and collect a number of essay-comment pairs to evaluate the effectiveness of this approach by the paradigm of zero-shot. The experimental results demonstrate the effectiveness of our approach, allowing the model to achieve higher performance in terms of correctness and diversity of comment generation.

In summary, our paper makes the following contributions:

- We propose a two-stage comment generation framework for the task of automated essay comment generation. It mimics the behaviour of students in the psychology of writing, which is beneficial in higher quality and diversity of the generated comments.
- We construct a knowledge base containing about five hundred writing cases of high-quality essays, and we collect over one hundred essay-comment pairs as our open source data.
- Our model improves significantly on both BLUE and Distinct scores, indicating the effectiveness of this framework.

2 Related Work

2.1 Automated Essay Evaluation

There have been wide explorations for automatic essay evaluation. Attali et al. [1] build an essay scoring system, known as the E-rater, which uses some hand-crafted features such as grammar. With the development of deep neural networks (DNN), several studies have utilized pre-trained language models (PLM) and well-designed algorithms to predict essay scores. LLM further advances the task of essay comment generation. The Linggle Write system [14] is capable of analyzing essays by focusing on grammatical and rhetorical perspectives. In a recent study, Zhang et al. [18] generate essay comments in the Chinese narrative essays, focusing on essay comments related to analysing rhetorical and descriptive writing skills.

2.2 LLM Prompting

Recently, the semantic understanding and reasoning capabilities of pre-trained language models have been significantly improved by scaling improvements with larger datasets and model sizes. The results show that large language models can be adapted to downstream tasks by inference alone [6], without parameter updates. Subsequent research has further improved the performance of in-context learning. Researchers have proposed advanced prompt formats from a variety of downstream tasks.

What's more, LLM has demonstrated strong ability in in-context learning and reasoning. With the few-shot prompting strategy, LLM is able to perform much better on the specific in-context learning and reasoning problems. Wei et al. [16] propose the Chain-of-Thought prompting, which appends several reasoning steps before the input question. To leverage the power of demonstrative examples and reduce manual effort, Zhang et al. [19] propose Auto-CoT, which automatically obtains k examples by clustering the given input and allows the model to generate rationales for the selected examples.

3 Method

The task of essay comment generation can be formulated as follows: given an essay $X = (x_1, x_2, ..., x_N)$ with N words, the method should generate a comment Y containing several words. In this work, we make a study of automated comment generation on English expository essays, in which the comments paraphrase how to improve the quality of essay on writing thought ideas.

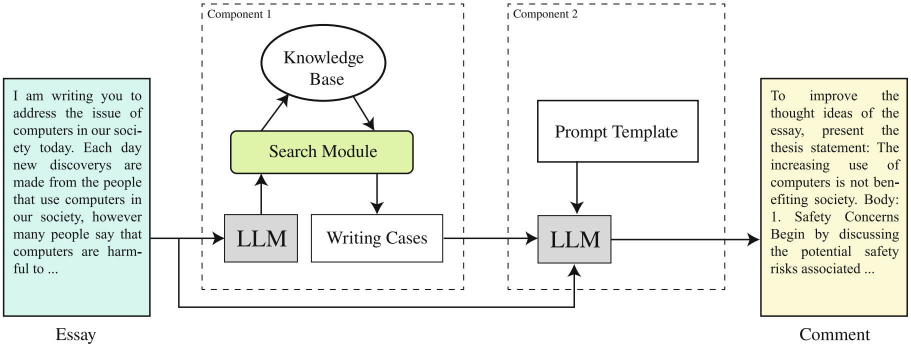

Component 1
Knowledge Base
Search Module
LLM
Writing Cases
Component 2
Prompt Template
LLM
I am writing you to address the issue of computers in our society today. Each day new discoverys are made from the people that use computers in our society, however many people say that computers are harmful to ...
Essay
To improve the thought ideas of the essay, present the thesis statement: The increasing use of computers is not benefiting society. Body: 1. Safety Concerns Begin by discussing the potential safety risks associated ...
Comment

Fig. 1. An overview of our framework. The two components correspond to the ordered generation stage.

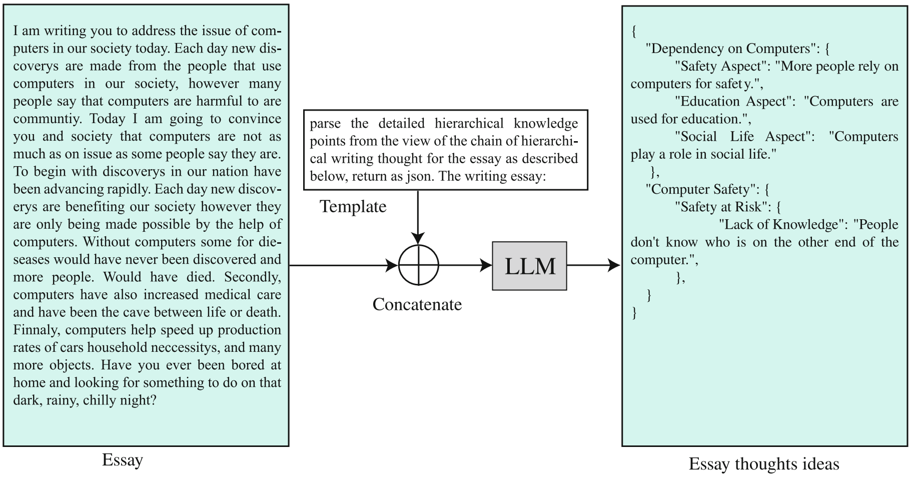
3.1 Comment Definition

Writing thought ideas in an essay are important indicators of the quality of essay. The comments generated by our method mainly contain the suggestions on how to improve the quality from the view of amplifying the writing thought ideas. Specifically, here are some of the elements to consider when manually evaluating essays: (1) Clarity of thought ideas. The ideas in an essay should present a clear, logical and well-structured line of thought. If not, we can suggest on how to amplify it by providing more explanations. (2) Coherence and cohesion. The ideas presented in the essay should be interconnected and form a coherent and cohesive narrative. If not, we can suggest on how to add or delete some ideas. (3) Evidence and examples. We can suggest that they include more examples to support their claims. Therefore, we aim to guide the model to generate more informative and diverse comments according to the above elements.

3.2 Two-Stage Comment Generation

As shown in Fig. 1, we propose a two-stage comment generation framework, in which the LLM first learns several examples of writing cases in excellent essays and then uses the template to improve the comment generation. Furthermore, in the processing of finding excellent writing cases, we first find the writing thought ideas within the input essay by parsing the essay though applying the LLM, then we design the specific search algorithm to match several examples in criteria of similar topics with the writing ideas. Moreover, in the comment generation stage, the well-designed template will integrate the previous cases and the input essay, which will be treated as a whole as the input to the LLM.

3.3 Writing Cases Searching



I am writing you to address the issue of computers in our society today. Each day new discoverys are made from the people that use computers in our society, however many people say that computers are harmful to are communtiy. Today I am going to convince you and society that computers are not as much as on issue as some people say they are. To begin with discoverys in our nation have been advancing rapidly. Each day new discoverys are benefiting our society however they are only being made possible by the help of computers. Without computers some for dieseases would have never been discovered and more people. Would have died. Secondly, computers have also increased medical care and have been the cave between life or death. Finnaly, computers help speed up production rates of cars household neccessitys, and many more objects. Have you ever been bored at home and looking for something to do on that dark, rainy, chilly night?

parse the detailed hierarchical knowledge points from the view of the chain of hierarchical writing thought for the essay as described below, return as json. The writing essay:

```
{
   "Dependency on Computers": {
         "Safety Aspect": "More people rely on computers for safet y.",
         "Education Aspect": "Computers are used for education.",
         "Social Life Aspect": "Computers play a role in social life."
    },
   "Computer Safety": {
         "Safety at Risk": {
                  "Lack of Knowledge": "People don't know who is on the other end of the computer.",
         },
   }
}
```

Fig. 2. The flow of parsing the given essay to the writing thought ideas. Taking advantage of the ability of LLM to follow instructions, we can directly use the LLM to generate the text containing the hierarchical writing thought ideas of the given essay, which can be easily extracted and saved in a *json* file. The writing thought idea consists of two parts, namely, the idea topic and the idea content. For example, the key in the *json* file is the idea topic, and the corresponding value is the idea content.

Parse Out the Writing Thought Ideas. Writing thought ideas are important to inspire and motivate writing and can assist in the use of attractive skills and diverse expression in writing. In essence, students tend to connect in the mind for the writing cases with related topics, learning their commonalities. Therefore, we first need a powerful tool to parse out the content of the input essay for related topics, which has good in-context learning capabilities to understand the textual content. Fortunately, the LLM itself has this capability. Hence, we utilize the LLM to parse out the essay, in which we can implement this step by $LLMprompting$. As shown in Fig. 2, we can see the flow of this parse and get part of the thought ideas of the essay. The $ideas$:

$$ideas = LLM(X, P) \tag{1}$$

where the X is the input essay and the P is the text template in parsing stage.

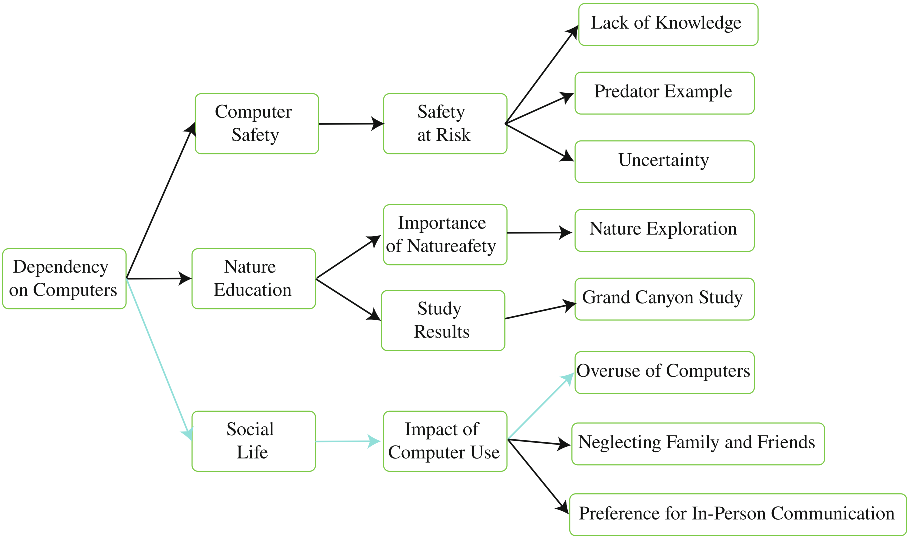


Fig. 3. An example of the hierarchical writing thought ideas in an excellent essay. As we can see from the figure, the root of the tree is the theme of the essay, then the non-leaf nodes are the topic of different part of the essay, and the leaf nodes are the concise content of the ideas in essay.

Knowledge Base. The challenge here is to express the structure of the thought ideas in an essay, which is hierarchical, structured and related between several writing thought ideas. Recent studies have shown that every document has its own structure [11,17], and the thought ideas in most essays can be illustrated through a tree structure. As shown in Fig. 3, each node in tree represents a thought idea in the essay, which has a lead from the previous idea and may branch to several other ideas, thus making the content of the whole essay grow up. Furthermore, we can see from the Fig. 3, the thought idea in the blue chain, called “Impact of Computer Use”, can be led from “Social Life” in the topic of “Dependency on Computers” and has some actual impact in real life. What’s more, because each essay can be transformed in this way, we can use this approach to build a content-rich knowledge base that includes several high-quality essays from a variety of topics, while using convenient data structures that help with subsequent searches. Intuitively, this type of data can be easily stored using JSON files.

Search Algorithm. The chain of writing thought ideas can provide students with the whole reasoning steps between several ideas. In this step, we need to search some excellent cases of the chain of writing thought ideas, since we have parsed out the writing thought ideas within the input essay. To do this, we design a searching-based algorithm, where the core idea is to find the $top - K$ chains containing the parsed writing ideas as much as possible. Specifically, every chain will be selected if itself contains the idea in the set of parsed writing thought ideas of the input essay. Then, we sort each chain according to the number of target ideas it contains, and finally select the $top - K$ chains. The pseudo-code is shown below:

Algorithm 1. Search Algorithm

```
Input: $K$, $ideas$, Knowledge Base, similarity threshold $t$
Output: the $top - K$ chains of the writing thought ideas $S$
Set total = {}
Set S={}
for $idea$ in $iedas$ do
    for every $tree$ in Knowledge Base do
        for every $node$ in $tree$ do
            if $node.idea$ is similar to $idea$ then
                total.append($node$)
                $p_{node}$ = find_parent_node($node$)
                $c_{node}$ = find_children_node($node$)
                $chain$ = construct_chain($p_{node}$, $node$, $c_{node}$)
                S.append($chain$)
            end if
        end for
    end for
end for
String $chain_{in}$ = construct_chain($ideas$)
S = filter_top_K($K$, $chain_{in}$, $S$)
```

Writing Case. The several cases of writing thought ideas can be used as evidences to inspire writing with authenticity, completeness, and diversity. We can use a formal format to illustrate this chain relationship, and the most basic method is to use arrows ↔ to represent progressive relationships. If extension is required, we can also design more specific relationship representations as needed. This demonstrates the adaptability and scalability of this method. As shown in the Table 1, these are some examples.

Table 1. Some examples of the case of the writing thought ideas. There is progressivity between nodes within each chain of the writing thought ideas, indicating the relevance, coherence and clarity in the contextual ideas.

Examples	
Id	Chain Content
1	Nature Education ↔ Importance of Nature safety ↔ Nature Exploration
2	Nature Education ↔ Study Results ↔ Grand Canyon Study
3	Social Life ↔ Impact of Computer Use ↔ Neglecting Family and Friends

3.4 Comment Generation

In the comment generation stage, we focus on how to guide LLM to learn some cases of writing thought ideas in order to improve the quality of the generated comments on the given essay. Specifically, a template have been carefully designed, which is used to integrate the writing thought cases and the essay. Finally, the whole text will be as an input to the LLM to generate comments on how to improve the essay. The *comments*:

$$comments = LLM(X, S, P) \tag{2}$$

where the X is the input essay, S is the writing cases from the output of the first component and the P is the text template in the comment generation stage.

4 Experiments

4.1 Dataset

Our task is to automatically generate comments for English expository essays on how to improve the quality of the essay, especially to amplify the writing thought ideas. As there is no dataset available for our task, we manually collect a English dataset to evaluate our model. Specifically, for the knowledge base, we collect about 500 expository essays with the corresponding writing thought ideas. For the evaluation, we collect over 100 expository essays with the corresponding annotated comment. For the process of annotation, we use the API of the ChatGLM-Pro model[1] to generate the preliminary comment and then perform manual secondary verification and supplementation, with the aim of obtaining comments as rich as possible.

[1] https://open.bigmodel.cn/.

4.2 Baseline

We use the llama2-13B of the chat version[2] as our backbone LLM, which is a collection of pre-trained and fine-tuned generative text models with 13B scale parameters and optimised for dialogue use cases. This LLM can directly generate comments on the input essay with the user query.

4.3 Experiment Settings

In the writing thought ideas parsing and comments generation stages, we use top-p sampling with $p = 0.95$ combined with beam search (number of beam is 1) while the parameter of *temperature* is 0.01. The fixed value of the parameter makes it easier for the LLM to generate a consistent context for multiple times. To improve the inference speed, we deploy our model on two GPU-3090Ti, the average cost of comment generation for an essay is 5.6 s.

4.4 Evaluation Metrics

The following automatic metrics will be adopt for the evaluation of comment generation.

BLEU (B-n). We use n = 1, 2 to evaluate the n-gram overlap between generated comments and ground truth comments [10]. BLUE is an algorithm for evaluating the quality of text between the output of a machine and that of a human.

Distinct (D-n). We use n = 2, 3, 4 to measure the generation diversity by the ratio of distinct n-grams to all the generated n-grams [7]. The distinct is a metric used to measure the degree of diversity in the content of machine-generated text.

4.5 Results

Table 2. Automatic evaluation results. All result values are multiplied by 100. The best performance is highlighted in **bold**.

Models	B-1	B-2	D-2	D-3	D-4
Llama2	35.39	11.65	10.74	29.76	43.86
Our Model	**42.58**	11.06	**11.47**	**31.61**	**46.13**
GroundTruth	100	100	16.25	41.18	58.75

Table 2 shows the automated evaluation results. Compared with the results of the **Llama2**, our method improves significantly on both BLUE and Distinct scores, indicating higher quality and diversity of the generated comments.

[2] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf.

4.6 Analysis

As for the comparison of the results, we can draw the following analysis: (1) For the fact that the BLUE scores are generally not high, intuitively because the given ground truth comments may not cover all aspects, so that some of the correct answers generated by the model are not reflected in the labelled comments. (2) The first component, which searches for some cases of the writing thought ideas in excellent essays, has a positive effect on diversity, suggesting that the model may tend to benefit from learning some examples and generate more diverse comments.

5 Examples

An example of the generated comment for the given essay is shown below.

The Essay:
I believe that computers are very helpful for many people by searching for information or talking to friends. Although many children should spend time outdoors, computers are an environment that takes up enough space to help children learn and interact with others. I think children should have the right to stay on the computer for as long as they want. Nevertheless, they should still make time to go out and play, play with friends, and so on. Computers are fun, but it's always important to go out and enjoy nature. Another reason why students should not spend all their time on computers is because of their health. Many children grow up to engage in computer production or programming. Some children receive education online or go to teach others. Computers are an uninterrupted foundation of intelligence that can help people learn, play, and discover. Many people dream of one day being able to build a computer or create new programs. All students should have the right to spend time on computers, but also the right to save time outdoors and spending time with friends. Everyone has their own ideas and whether staying on the computer all day is a good thing. I think children should not do this because it may affect their health and the opinions of others towards them.

Search Output:
Example 1: computers are helpful $\leftrightarrow$ learning and social interaction $\leftrightarrow$ balanced approach; Example 2: programming and design $\leftrightarrow$ education online $\leftrightarrow$ education and career development; Example 3: time cost $\leftrightarrow$ physical and mental health $\leftrightarrow$ balance between screen and activities

Comment:
Here are some suggestions for improving the essay you provided.
Thesis Statement: Start your essay with a clear and concise thesis statement that reflects your main argument. In your case, it might be something like: "While computers offer numerous benefits, it is essential for kids to strike a balance between screen time and outdoor activities to ensure a healthy and well-rounded upbringing." Introduction: Begin with an engaging hook that grabs the reader's

attention. It could be a relevant statistic, a question, or a thought-provoking statement about the influence of computers on children. Introduce the topic and your thesis statement.
Body Paragraphs: Provide evidence and examples to support your claims. For instance, you can discuss the benefits of computer use for learning and interaction with peers.

6 Conclusion

We present a two-stage comment generation framework, where the LLM first learns several examples of excellent essay writing cases and then uses the template to improve the comment generation. Furthermore, when searching for writing cases, we start by parsing out the ideas of the input essay. Then, we design a search algorithm to find several cases of the chain of the writing ideas. In the comment generation stage, we use a carefully designed template that combines the previous cases and the input essay and treats it as input to LLM to generate comments on how to improve the quality of the essay from the perspective of extending contextual ideas. We manually collect some English expository essays to construct the knowledge base and some essay-comment pairs. Extensive experiments demonstrate the effectiveness of our method in the real application. In future work, we will add more technology to make the performance better. We expect that our work will contribute to further research on this application and research and benefit both teachers and students.

Acknowledgement. This work is supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_1775) under Grant 181200003023202. The project name is the "Research and Implementation of Automated Essay Scoring".

References

1. Attali, Y., Burstein, J.: Automated essay scoring with e-rater® v. 2. J. Technol. Learn. Assess. **4**(3) (2006)
2. Chowdhery, A., et al.: Palm: scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
3. Deumert, A.: Mimesis and mimicry in language–creativity and aesthetics as the performance of (dis-) semblances (2018)
4. Guéguen, N., Martin, A., Meineri, S.: Mimicry and helping behavior: an evaluation of mimicry on explicit helping request. J. Soc. Psychol. **151**(1), 1–4 (2011)
5. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366 (2021)
6. Houlsby, N., et al.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, pp. 2790–2799. PMLR (2019)
7. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)

8. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
9. Manakul, P., Liusie, A., Gales, M.J.: Selfcheckgpt: zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896 (2023)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BleU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
11. Power, R., Scott, D., Bouayad-Agha, N.: Document structure. Comput. Linguist. **29**(2), 211–260 (2003)
12. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. arXiv preprint arXiv:1809.02156 (2018)
13. Thoppilan, R., et al.: Lamda: language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
14. Tsai, C.T., Chen, J.J., Yang, C.Y., Chang, J.S.: Lingglewrite: a coaching system for essay writing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 127–133 (2020)
15. Wang, B., et al.: Towards understanding chain-of-thought prompting: an empirical study of what matters. arXiv preprint arXiv:2212.10001 (2022)
16. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837 (2022)
17. Zhang, Y., Yu, X., Cui, Z., Wu, S., Wen, Z., Wang, L.: Every document owns its structure: inductive text classification via graph neural networks. arXiv preprint arXiv:2004.13826 (2020)
18. Zhang, Z., Guan, J., Xu, G., Tian, Y., Huang, M.: Automatic comment generation for Chinese student narrative essays. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 214–223 (2022)
19. Zhang, Z., Zhang, A., Li, M., Smola, A.: Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493 (2022)