# Guiding Large Language Models via External Attention Prompting for Scientific Extreme Summarization

**Yuan Chang, Ziyue Li, Xiaoqiu Le**[*]
National Science Library, Chinese Academy of Sciences
Department of Information Resources Management, School of Economics and Management,
University of Chinese Academy of Sciences
{changyuan, liziyue, lexq}@mail.las.ac.cn

## Abstract

Scientific extreme summarization, the task of generating concise one-sentence summaries (TLDRs) for scientific papers, presents significant challenges due to the need for deep domain-specific understanding and the ability to distill salient information. This study identifies the critical role of titles and keywords in enhancing TLDR generation through quantitative analysis. We propose a novel method, External Attention Prompting (EAP), which leverages LLMs by guiding them to focus on the most critical parts of the source text through varying degrees of attention signals. Our method employs Markdown emphasis syntax to annotate attention levels, enabling LLMs to prioritize salient information effectively. Extensive experiments demonstrate that EAP significantly outperforms baseline methods across various LLMs and metrics in both zero-shot and few-shot settings. Further evaluations by GPT-4 demonstrate that EAP can enable LLMs to generate TLDRs of higher human-aligned quality.

## 1 Introduction

The rapid growth of scientific literature has made it increasingly difficult for researchers to keep up with the research frontiers and quickly identify the most relevant and impactful information. In this context, recent work (Cachola et al., 2020; Lu et al., 2020; Mao et al., 2022; Takeshita et al., 2022; Atri et al., 2023; Stiglic et al., 2023; Syed et al., 2024) has studied the problem of scientific extreme summarization, which involves generating a concise one-sentence summary (TLDR: Too Long; Didn't Read) that captures the key aspects of a scientific paper. Scientific extreme summarization is a challenging task that requires a deep understanding of domain-specific content to accurately identify and distill the most salient points (Cachola et al., 2020). This task demands advanced methods capable of leveraging domain knowledge and prioritizing the most important content.

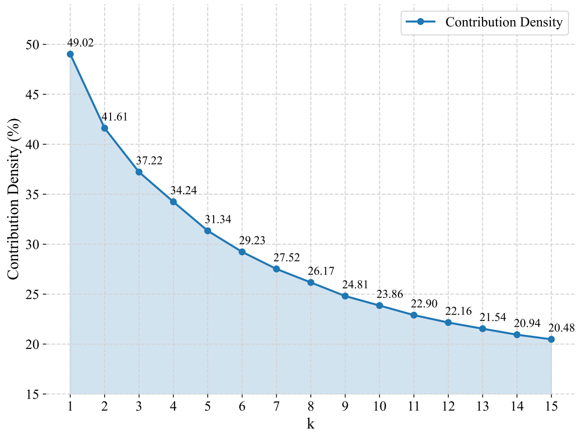

Figure 1: Contribution density of top-k TextRank keywords from abstracts on the SciTLDR training set, showing diminishing returns as k increases.

Previous studies have found that key content such as the titles and keywords of source documents are helpful to improve performance in summarization task, as they often contain key information (Chen et al., 2019; Çano and Bojar, 2019; Lee et al., 2020; Cachola et al., 2020; Koto et al., 2022; Mao et al., 2022). Our analysis on the SciTLDR dataset (Cachola et al., 2020) further indicates that their contributions to generating highly compressed summaries differ as they contain diverse density of key information (see Section 2), suggesting that they should be given unequal levels of importance.

In this work, we leverage large language models (LLMs) to tackle the scientific extreme summarization task, as they have demonstrated emergent abilities such as in-context learning and made remarkable advances in solving various NLP problems (Wei et al., 2022; Zhao et al., 2023). We propose a novel method called External Attention Prompting (EAP) that enhances the performance of LLMs by guiding them to pay varying degrees of attention to

[*]Corresponding Author

*Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 226–242
August 16, 2024 ©2024 Association for Computational Linguistics

different segments of the source text. Our method involves constructing prompts that assign different levels of attention signals to the n-grams in the source text based on their importance, using Markdown emphasis syntax to annotate the attention strength. This approach enables the LLMs to prioritize the salient information and generate more accurate and informative TLDRs. Experimental results demonstrate that our proposed prompting method significantly enhances the performance of LLMs in this task, outperforming baseline methods.

The main contributions of this work are as follows:

- We provide insights into the importance of different source input components (title, abstract, and keywords) for scientific extreme summarization, highlighting the role of the title and keywords in capturing key information.
- We propose a novel prompting method, External Attention Prompting, that enhances the performance of LLMs in scientific extreme summarization by guiding them to focus on the most critical parts of the source text.
- We conduct extensive experiments on the SciTLDR dataset, demonstrating the effectiveness of our approach in improving the quality of generated TLDRs.

## 2 Source Contribution Analysis

Generating high-quality TLDRs requires capturing precisely the salient aspects of the document. In this section, we conduct a preliminary experiment to investigate whether components such as the title, abstract and keywords contribute differently to the formation of a TLDR. By identifying the most influential parts, we aim to enhance the summarization process by directing the model's attention to the most critical information.

To do this, we firstly introduce the concept of contribution density, which measures the averaged contribution of tokens from a specific source input components (e.g., title, abstract, and keywords) to the gold TLDR. We first tokenize the input components and remove stopwords. Then, the contribution density $CD(c)$ for an input component $c$ is calculated as:

$$CD(c) = \frac{\sum_{t \in c} TF(t, TLDR)}{|c|} \tag{1}$$

where $t$ is a token in the input component $c$, $TF(t, TLDR)$ is the term frequency of token $t$ in the gold TLDR, and $|c|$ is the total number of tokens in the input component $c$.

The calculation of contribution density draws inspiration from methods used in previous works (See et al., 2017; Narayan et al., 2018; Cachola et al., 2020) to assess abstractiveness and extractiveness of generated summaries. However, in this context, it serves as an analytical tool rather than a direct evaluation metric for summary quality assessment. The term "contribution" represents how frequently the words from source input appear in the gold TLDR. We use contribution density as a proxy to quantify the relative importance of different source input components in contributing to extreme summarization.

We take the training set of SciTLDR (Cachola et al., 2020) for analysis, and calculate the contribution density for title and abstract, which are 49.2 and 11.9, respectively. The results align with our intuition that the title often captures the most core idea of a paper. While titles may contain less overall information than abstracts, the information they do contain is more densely packed with content relevant to the TLDR. In other words, any given n-gram from the title is more likely to contribute to the TLDR than a random n-gram from the abstract. Abstracts may contain more information that is either less important or not sufficiently crucial to be included in the TLDR.

Additionally, we calculate the contribution densities for the top-k keywords from abstract using TextRank algorithm (Mihalcea and Tarau, 2004), the results are shown in Figure 1. The contribution density decreases as k increases, indicating diminishing returns with more keywords. This highlights the importance of selecting a concise set of highly relevant keywords.

The above findings from our source contribution analysis underscore the necessity of focusing on the most important parts of the input text when generating a TLDR.

## 3 External Attention Prompting

Based on the insight from section 2, we propose a novel method called **External Attention Prompting (EAP)**. The core idea of our method is to guide LLMs to focus on the most critical parts of the source text when generating TLDRs. As illustrated in Figure 2, our method leverages external signals

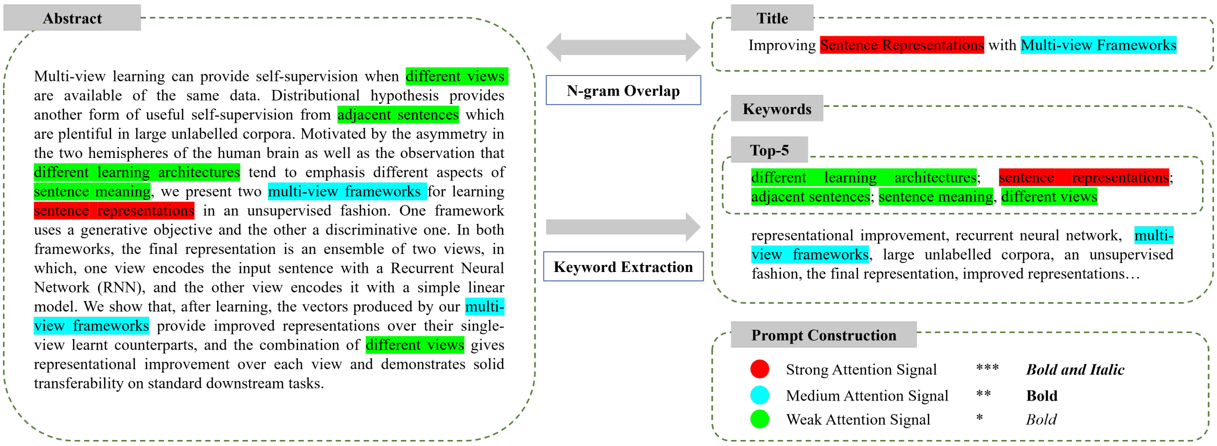
Abstract
Multi-view learning can provide self-supervision when different views are available of the same data. Distributional hypothesis provides another form of useful self-supervision from adjacent sentences which are plentiful in large unlabelled corpora. Motivated by the asymmetry in the two hemispheres of the human brain as well as the observation that different learning architectures tend to emphasis different aspects of sentence meaning, we present two multi-view frameworks for learning sentence representations in an unsupervised fashion. One framework uses a generative objective and the other a discriminative one. In both frameworks, the final representation is an ensemble of two views, in which, one view encodes the input sentence with a Recurrent Neural Network (RNN), and the other view encodes it with a simple linear model. We show that, after learning, the vectors produced by our multi-view frameworks provide improved representations over their single-view learnt counterparts, and the combination of different views gives representational improvement over each view and demonstrates solid transferability on standard downstream tasks.
N-gram Overlap
Keyword Extraction
Title
Improving Sentence Representations with Multi-view Frameworks
Keywords
Top-5
different learning architectures; sentence representations; adjacent sentences; sentence meaning, different views
representational improvement, recurrent neural network, multi-view frameworks, large unlabelled corpora, an unsupervised fashion, the final representation, improved representations…
Prompt Construction
Strong Attention Signal *** Bold and Italic
Medium Attention Signal ** Bold
Weak Attention Signal * Bold

Figure 2: Overview of the proposed External Attention Prompting (EAP) method.

derived from the paper's title and keywords to modulate the attention of the LLM. By emphasizing different segments of the source text with varying degrees of importance, we aim to enhance LLMs' ability to distill the essence of the scientific paper into a concise and informative TLDR.

### 3.1 Attention Signal Acquisition

To effectively guide the model's attention, this method first acquires attention signals. We hypothesize that certain n-grams within the text, particularly those overlapping with the title and keywords, are more indicative of the core content. To identify these n-grams, we perform the following steps:

**Title Overlap Extraction** We extract n-grams from the title and identify their occurrences in the source text. To ensure relevance, we remove stop words and retain only the longest overlapping n-grams.

**Keyword Extraction** Using the TextRank algorithm (Mihalcea and Tarau, 2004), we extract keywords from the source text, removing stop words to focus on meaningful terms.

**Attention Signal Levels Definition** We define attention signals of different levels as follows:

- **Strong Attention Signal**: N-grams present in both the title and the top-5 keywords.
- **Medium Attention Signal**: N-grams present in both the title and any keyword.
- **Weak Attention Signal**: N-grams present in either the title or the top-5 keywords.

### 3.2 Prompt Construction

To operationalize these attention signals, we employ Markdown emphasis syntax to "visually" highlight the importance of different n-grams within the source text. Markdown is chosen for its inherent three-level emphasis system (italic, bold, and bold-italic) and because there is a large amount of Markdown-formatted text in the training corpus of LLMs, enabling them to understand the emphasis cues implied by such annotations. Specifically, we mark n-grams of strong attention signal with *** (***bold and italic***), n-grams of medium attention signal with ** (**bold**) and n-grams of weak attention signal with * (*italic*).

By using Markdown's built-in emphasis features, we avoid the need for custom formatting, allowing the model to leverage its pre-existing understanding of these conventions. This approach ensures that the model can efficiently recognize and prioritize the marked text without additional fine-tuning or instruction.

The marked source text, along with necessary instructions, is then used to construct the prompt, which is fed into the LLMs to generate the TLDR. Furthermore, we explicitly instruct the model to pay special attention to the emphasized parts of the text. The prompt template used for this method can be seen in Appendix B.

The term "External" in this method reflects two key aspects:

**External attention signals.** The attention signals are derived from external information such as the title and keywords, rather than being internally recognized by the model. This external guidance directs the model to follow specific cues.

**External implementation.** The attention mechanism is implemented through external input (i.e., the prompt) without fine-tuning LLMs. This makes the method applicable even when model parameters are inaccessible.

The proposed method is highly generalizable and can be applied to various tasks requiring controlled text generation or tasks where the importance of different parts of the input is known a priori. The attention signals can be provided by domain experts, making the approach adaptable to different contexts. This method is particularly suitable for tasks with long input texts, such as summarization and reading comprehension, where the model needs to focus on key content.

The external attention mechanism makes use of the LLMs' capabilities in in-context learning and instruction following. By providing clear, externally defined attention signals, LLMs can better understand and prioritize the critical content, leading to improved performance in specific tasks.

## 4 Experiments

### 4.1 Experimental Setup

**Dataset.** We evaluate our method on SciTLDR (Cachola et al., 2020), a widely used benchmark for scientific extreme summarization, featuring high-quality, human-generated gold TLDRs. Our experiments focus on the abstract-only subset (including title), rather than the AIC (abstract+introduction+conclusion) and full text subsets, to reduce API usage costs and to accommodate the limited context length of most LLMs. Additionally, abstracts are generally more accessible than full texts , making this setting more applicable to real-world scenarios.

**Baselines.** We compare the proposed EAP method with two baselines:

- **Standard Prompting (SP)**: This baseline only provide necessary instructions and the abstracts to LLMs to generate the TLDRs. No additional guidance or emphasis is included.

- **Explicit Guidance Prompting (EGP)**: In this baseline, the abstract is supplemented with the paper's title and the top-5 keywords extracted using the TextRank algorithm. The LLMs are explicitly instructed to "pay attention to" the title and keywords, thereby directing its focus towards these critical components. This method provides explicit guidance to the model without altering the abstract.

- **Random Attention Prompting (RAP)**: This baseline is a variant of our proposed EAP method. It randomly applies different levels of emphasis to n-grams in the source text, maintaining the same number of emphasized n-grams and n-gram lengths as the corresponding EAP method. This baseline is designed to investigate whether the performance changes in EAP are due to the specific attention signals or simply the presence of Markdown emphasis, regardless of the content being emphasized.

All methods are evaluated under both zero-shot and few-shot settings. In the few-shot setting, we use the same two demonstration examples in prompt for all methods, with the input processed according to each respective method. The specific prompts utilized for all methods are detailed in Appendix B.

**Large language models.** To evaluate the effectiveness and generalizability of our proposed approach, we conduct experiments leveraging a diverse set of LLMs, including GPT-3.5, Llama-3, GLM-4, and DeepSeek-V2, encompassing both open-source and black-box models. The detailed information on the versions of LLMs used in our experiments is summarized in Table 2.

**Evaluation metrics.** In our evaluation, we use Rouge-1, Rouge-2, and Rouge-L metrics (Lin, 2004), consistent with previous work in summarization (Narayan et al., 2018; Lewis et al., 2020; Cachola et al., 2020). Considering SciTLDR provides multiple target summaries for a given paper, we follow Cachola et al. (2020) to taking the maximum Rouge score calculated over multiple gold TLDRs as the final Rouge score. This can effectively handles the variability in gold TLDRs.

Additionally, we use the Rouge-K metric (Takeshita et al., 2024), a keyword-oriented evaluation metric which specifically assesses the inclusion of essential keywords in the generated summaries, as it aligns with our goal of ensuring that the generated TLDRs capture the most critical content from the source documents.

**Implementation details.** All LLMs utilized in this work are accessed via API calls and configured with the same generation parameters: temperature of 0.1 and a maximum tokens limit of 128. Evaluation results for each LLM are averaged over three independent inferences for each input to mitigate the impact of sampling randomness of LLMs.

| | Method | GPT-3.5 | | | Llama-3 | | | GLM-4 | | | DeepSeek-V2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | R-1 | R-2 | R-L | R-1 | R-2 | R-L | R-1 | R-2 | R-L |
| Zero-shot | SP | 32.04 | 12.04 | 26.04 | 32.54 | 12.36 | **26.92** | 31.46 | 10.73 | **26.00** | 30.49 | 10.48 | 25.09 |
| | EGP | 32.56 | 12.40 | **26.48** | 32.31 | 11.96 | 26.50 | 31.13 | 10.20 | 25.39 | 30.41 | 10.61 | 24.98 |
| | RAP | 31.26 | 11.19 | 25.36 | 32.15 | 11.95 | 26.22 | 31.12 | 10.40 | 25.61 | 29.42 | 9.72 | 23.72 |
| | EAP* | **32.64**† | **12.50**† | 26.43 | **32.74** | **12.37** | 26.78 | **31.54** | **10.92** | **26.00** | **30.97** | **10.91** | **25.42** |
| Few-shot | SP | 33.76 | 13.22 | 27.91 | 34.29 | 13.24 | 28.33 | 31.76 | 10.90 | 26.27 | 33.73 | 13.03 | 28.06 |
| | EGP | 33.32 | 12.90 | 27.40 | 33.76 | 12.90 | 27.75 | 31.79 | 11.03 | 26.05 | 33.62 | 12.76 | 28.17 |
| | RAP | 33.01 | 12.77 | 27.23 | 34.26 | 13.28 | 28.25 | 31.86 | 10.94 | 26.27 | 32.34 | 11.82 | 26.88 |
| | EAP* | **34.39**† | **13.69**† | **28.49**† | **34.78**† | **13.80**† | **28.82**† | **32.34**† | **11.39**† | **26.84**† | **34.23**† | **13.34** | **28.58** |

Table 1: Rouge scores for different prompting methods across various LLMs in zero-shot and few-shot settings. Best results per metric are **bolded**. The method marked with * is the proposed method in this paper. Scores marked with † indicate that the EAP method significantly ($\alpha = 0.05$) outperforms the Standard Prompting (SP) method, determined using the Wilcoxon signed-rank test (Dror et al., 2018).

| LLM | Version |
|---|---|
| GPT-3.5 | gpt-3.5-turbo-0125 |
| Llama-3 | llama-3-70b-instruct |
| GLM-4 | glm-4 (2024-01-16) |
| DeepSeek-V2 | deepseek-chat (2024-05-17) |

Table 2: Specifications of large language models used in the experiments.

### 4.2 Main Results and Analysis

The experimental results, as presented in Table 1, demonstrate the effectiveness of the proposed EAP method across various LLMs and metrics in both zero-shot and few-shot settings. This indicates that the external attention signals effectively guide the LLMs to generate more accurate and informative TLDRs.

**Zero-shot results.** In the zero-shot setting, EAP outperforms all baseline methods (SP, EGP and RAP) across all LLMs on Rouge-1/2, suggesting that the external attention signals provided by EAP are effective even without prior examples. However, in the case of the Rouge-L metric, the EAP method underperforms compared to the baseline methods for GPT-3.5 and Llama-3, possibly due to the additional complexity and noise introduced by this attention mechanism.

**Few-shot results.** In the few-shot setting, EAP consistently outperforms all baseline methods across all metrics and LLMs. The majority of these improvements are statistically significant, as evidenced by the Wilcoxon signed-rank test ($\alpha = 0.05$). This indicates that providing demonstration examples helps LLMs learn how to effectively distill the core content by focusing on segments of varying importance, leading to the generation of more precise and comprehensive TLDRs. This enhanced performance can be attributed to the strong in-context learning capabilities of LLMs.

**Explicit vs. implicit guidance.** The EGP method, which explicitly provides titles and keywords as additional input, shows mixed results. In some cases, it performs worse than the SP method, which does not include any additional guidance. This suggests that merely adding extra information without fine-grained control can confuse the model, leading to inconsistent results. In contrast, the EAP method implicitly emphasizes different parts of the source text, allowing the model to better understand and prioritize the critical content. This implicit approach appears to be more effective in guiding the model's attention, resulting in more stable and improved performance.

**Random Attention Prompting results.** The RAP baseline generally underperforms compared to SP and EAP methods across all LLMs and metrics in both zero-shot and few-shot settings. In the zero-shot setting, RAP often performs worse than SP, suggesting that random emphasis can be detrimental to the LLMs' understanding of the source text. In the few-shot setting, while RAP shows some improvement over SP in certain cases (e.g., for GLM-4), it still consistently underperforms compared to EAP. These results indicate that the performance gains of EAP are not merely due to the presence of Markdown emphasis, but rather the strategic placement of attention signals on important content. The effectiveness of EAP over RAP underscores the importance of deriving attention signals from key content such as titles and keywords, rather than applying emphasis randomly.

**Results across LLMs.** The results indicate that the EAP method consistently enhances performance across different models, demonstrating its generalizability and effectiveness. However, the degree of improvement varies among models, which may be related to the inherent capabilities of each model. For instance, models with stronger in-context learning abilities might benefit more from the EAP method, as they can better utilize the implicit attention signals to generate high-quality summaries.

**Rouge-K results.** The Rouge-K metric (Takeshita et al., 2024), which assesses the inclusion of essential keywords in the generated summaries, provides additional insights into the performance of the EAP method. As shown in Figure 3, in the zero-shot setting, EAP achieves higher Rouge-K scores compared to the baseline methods on Llama-3 and DeepSeek-V2. However, the EGP method surpasses EAP on GPT-3.5 and GLM-4, suggesting that explicitly providing keywords may be more effective for certain models in the absence of demonstration examples. In the few-shot setting, EAP consistently outperforms all the baseline methods across all models in terms of Rouge-K. This indicates that the combination of implicit attention signals and demonstration examples enables the models to better capture and include the most critical points in the generated TLDRs. The RAP method shows inconsistent performance on Rouge-K across different settings and models. In some cases, RAP outperforms SP or EGP methods, while in others, it underperforms. This variability can be attributed to the random nature of the attention signals in RAP, which may occasionally emphasize important content by chance. It is worth noting that the Rouge-K scores for all models and most methods are significantly lower in the few-shot setting compared to the zero-shot setting. This unexpected result may be attributed to the demonstration examples potentially biasing the models towards generating summaries that prioritize overall content coverage rather than focusing on specific keywords. Further investigation is needed to fully understand this phenomenon and its implications for keyword-oriented summarization tasks.

### 4.3 LLM as Evaluator

**Setup.** In addition to using traditional automatic metrics, we explore the use of LLMs as evaluators to assess the quality of generated TLDRs. This approach aims to address the limitations of automatic metrics, which may not fully capture the nuances of human preferences and the overall quality of the summaries (Zheng et al., 2023). By leveraging LLMs, we seek to provide a more comprehensive and human-aligned evaluation of the generated TLDRs.

To this end, we employ GPT-4 (version gpt-4-0613), which is currently recognized as one of the most advanced LLMs, to assess the quality of TLDRs generated by our method compared to a baseline method.

We randomly sample 100 data points and collect TLDRs generated by GPT-3.5 using both the baseline SP method and our EAP method under a few-shot setting. The evaluation criteria focus on the TLDR's ability to concisely capture the key aspects of a scientific paper while maintaining faithfulness to the source. To mitigate position bias (Zheng et al., 2023), each pair is evaluated twice with the positions of the two TLDRs swapped. A method is considered the winner for a sample only if both evaluations favored the same TLDR. In cases of conflicting results or ties, the sample is marked as a tie. The above process is repeated with three different random seeds, resulting in three rounds of evaluation.

**Results.** The results of GPT-4 evaluation are summarized in Table 3, showing that the proposed EAP method consistently outperforms the baseline SP method, achieving an average win rate of 37.33 % compared to the baseline's 26.00 %. The average tie rate is 36.67 %, with an average conflict rate of 23.67 %. These results indicate that the EAP method significantly enhances the performance of LLMs in generating TLDRs, as judged by GPT-4.

### 4.4 Ablation Study

To investigate the effects of different components in EAP method, we conduct an ablation study to address two key research questions: (1) Does the multi-level attention signal mechanism contribute to the performance improvement compared to using a uniform attention signal? (2) Are attention signals derived from the title essential for enhancing the

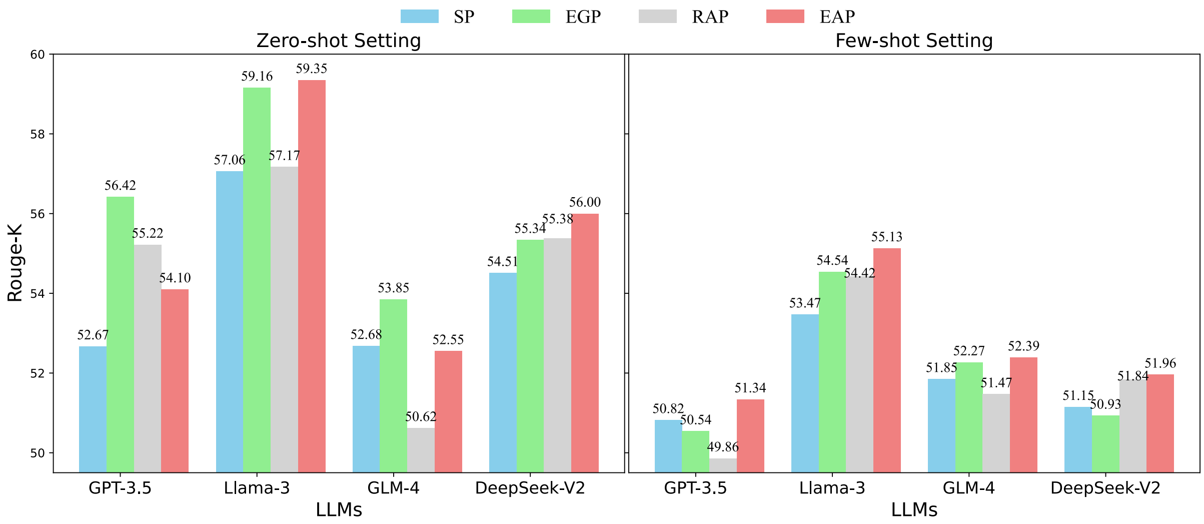

Figure 3: Rouge-K scores of SP, EGP, RAP and EAP methods across the tested LLMs.

| Rounds | SP Win Rate | EAP Win Rate | Tie (Conflict) |
|---|---|---|---|
| Round1 | 26% | 35% | 39% (27%) |
| Round2 | 23% | 40% | 37% (23%) |
| Round3 | 29% | 37% | 34% (21%) |
| Avg. | 26% | **37.33%** | 36.67% (23.67%) |

Table 3: GPT-4 evaluation results comparing EAP and SP methods for TLDR generation.

summarization quality?

To investigate the role of multi-level attention signals, we design an ablation variant of our method, in which we remove the differentiation between strong, medium, and weak attention signals. Instead, we uniformly apply a single level of attention signal by marking all relevant n-grams from the title and keywords with bold Markdown syntax.

To assess the importance of attention signals derived from the title, we create another ablation variant in which we exclude attention signals derived from the title and only mark the top-5 keywords extracted using the TextRank algorithm in the source text.

We conduct experiments using the same setup as described in Section 4.1, maintaining consistency in dataset, baselines, and evaluation metrics. The results are summarized in Table 4.

**Multi-Level Attention Signals** The variant without multi-level attention signals exhibit a slight decline in performance across vast majority metrics and LLMs compared to the original EAP method. This suggests that the differentiation between strong, medium, and weak attention signals plays a crucial role in guiding the LLMs to prioritize the most salient information effectively.

**Title-Derived Attention Signals**: The variant without title attention shows a noticeable reduction in performance, indicating that while the multi-level attention mechanism is advantageous, the title-derived signals play a more crucial role in capturing the core idea of scientific papers, thereby contributing significantly to the overall performance improvement.

## 5 Related Work

**Scientific Extreme Summarization** Scientific extreme summarization, the task of generating extreme one-sentence summaries (TLDRs) for scientific papers, has gained increasing attention in recent years. Several datasets have been proposed to facilitate research in this area, including SciTLDR (Cachola et al., 2020), a high-quality dataset with human-written summaries, and CiteSum (Mao et al., 2022), which contains automatically generated summaries based on citations. Additionally, multi-document(Lu et al., 2020), multilingual (Takeshita et al., 2022) and multimodal (Atri et al., 2023) datasets have been introduced to explore the task in diverse settings. Existing ap-

| | Method | GPT-3.5 | | | Llama-3 | | | GLM-4 | | | DeepSeek-V2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | R-1 | R-2 | R-L | R-1 | R-2 | R-L | R-1 | R-2 | R-L |
| Zero-shot | EAP | 32.64 | 12.50 | 26.43 | 32.74 | 12.37 | 26.78 | 31.54 | 10.92 | 26.00 | 30.97 | 10.91 | 25.42 |
| | w/o MLA | 32.41 | 12.19 | 26.20 | 32.67 | 12.41 | 26.68 | 31.49 | 10.81 | 25.82 | 30.97 | 10.91 | 25.39 |
| | Δ | -0.23 | -0.31 | -0.23 | -0.07 | +0.04 | -0.10 | -0.05 | -0.11 | -0.18 | 0.00 | 0.00 | -0.03 |
| | w/o TA | 31.61 | 11.50 | 25.65 | 32.30 | 12.06 | 26.34 | 31.22 | 10.60 | 25.70 | 30.89 | 10.62 | 25.19 |
| | Δ | -1.03 | -1.00 | -0.78 | -0.44 | -0.31 | -0.44 | -0.32 | -0.32 | -0.30 | -0.08 | -0.29 | -0.23 |
| Few-shot | EAP | 34.39 | 13.69 | 28.49 | 34.78 | 13.80 | 28.82 | 32.34 | 11.39 | 26.84 | 34.23 | 13.34 | 28.58 |
| | w/o MLA | 34.02 | 13.35 | 28.04 | 34.58 | 13.63 | 28.66 | 32.45 | 11.38 | 26.71 | 34.15 | 13.19 | 28.48 |
| | Δ | -0.37 | -0.34 | -0.45 | -0.20 | -0.17 | -0.16 | +0.11 | -0.01 | -0.13 | -0.08 | -0.15 | -0.10 |
| | w/o TA | 33.26 | 12.75 | 27.45 | 34.11 | 13.20 | 28.14 | 31.91 | 11.10 | 26.34 | 33.94 | 12.98 | 28.19 |
| | Δ | -1.13 | -0.94 | -1.04 | -0.67 | -0.60 | -0.68 | -0.43 | -0.29 | -0.50 | -0.29 | -0.36 | -0.39 |

Table 4: Ablation study results. "w/o MLA" denotes the variant without multi-level attention signals, and "w/o TA" represents the variant without title-derived attention signals. Δ indicates the performance difference between each variant and the full EAP method. The results demonstrate that both multi-level attention and title-derived signals contribute to the effectiveness of EAP, with title-derived signals playing a more crucial role.

proaches to scientific extreme summarization primarily rely on fine-tuned Transformer-based models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2023), which have achieved promising results. However, the potential of LLMs for this task remains largely unexplored. While LLMs are trained for general-purpose tasks and may not match the performance of task-specific fine-tuned models (Takeshita et al., 2024), their extensive domain knowledge and strong in-context learning abilities make them a promising avenue for investigation in scientific extreme summarization.

**Key Content Utilization for Summarization** Previous studies have explored various key content, such as titles, keywords, topics, and key segments, to enhance summarization performance (Chen et al., 2019; Koto et al., 2022; Srivastava et al., 2023; Tang et al., 2023). These methods typically focus on incorporating key content as additional input features or using the prediction of key content as an auxiliary task. For instance, Koto et al. (2022) investigated the impact of incorporating titles and keyphrases as structured input and target outputs in news summarization. Cachola et al. (2020) leveraging titles as auxiliary training signals to improve model performance in the scientific extreme summarization task. In this work, we use key content to compute attention signals, rather than as additional input or target, to guide LLMs to focus on the important parts of the source text.

**Prompt Optimization and Engineering** Prompt engineering has emerged as a crucial area of research to enhance the capabilities of LLMs on complex tasks. Early approaches focused on tuning soft prompts, which are continuous embedding vectors optimized using gradient descent methods (Li and Liang, 2021; Vu et al., 2022; An et al., 2022; Tam et al., 2023). However, these approaches require access to the Internal parameters of LLMs, limiting their applicability to black-box models. Another direction involves designing task-specific natural language instructions and selecting appropriate in-context demonstrations (Shin et al., 2020; Brown et al., 2020; Deng et al., 2022; Zhou et al., 2023; Sun et al., 2023). Despite the promising results, these methods may struggle to capture fine-grained, instance-specific (Li et al., 2023). proposes a method that generates instance-specific prompts tailored to each input using a small, tunable policy model optimized through supervised fine-tuning and reinforcement learning. In this work, we aim to provide LLMs with instance-specific attention signals derived from key content to enhance their performance in the task of scientific extreme summarization.

## 6 Conclusion

In this work, we have presented External Attention Prompting (EAP), a novel method for improving the performance of large language models (LLMs) in the task of scientific extreme summarization. By utilizing external attention signals derived from the title and keywords of scientific papers, EAP effectively guides LLMs to focus on the most salient information, resulting in more accurate and informative TLDRs. Our extensive experiments on the SciTLDR dataset demonstrate the superiority of EAP over standard prompting and explicit guidance methods, particularly in few-shot settings. The results underscore the importance of multi-level attention and the significant role of title emphasis in enhancing summarization quality. Future work could explore the application of EAP to other domains and tasks, further validating its versatility and effectiveness in various text generation scenarios.

## 7 Limitations

This work has the following limitations: (1) The heuristic-based design and computation of attention signals in EAP may not fully capture salient information in diverse contexts. Future work could explore automated methods. (2) The keyword extraction algorithm used in this work for computing attention signals only considers TextRank. Future research should explore and compare the effectiveness of various keyword extraction methods. (3) Our experiments focus on the abstract-only subset of SciTLDR, which may not fully represent the challenges of summarizing entire scientific papers. Furthermore, SciTLDR is focused on the computer science domain, which may limit the generalizability of our findings.

## Acknowledgments

The authors thank all the anonymous reviewers for their valuable comments and constructive feedback. The authors acknowledge financial support from the National Social Science Fund of China (No. 23BTQ102). Xiaoqiu Le is the corresponding author.

## References

Shengnan An, Yifei Li, Zeqi Lin, Qian Liu, Bei Chen, Qiang Fu, Weizhu Chen, Nanning Zheng, and Jian-Guang Lou. 2022. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models. *Preprint*, arXiv:2203.03131.

Yash Kumar Atri, Vikram Goyal, and Tanmoy Chakraborty. 2023. Fusing multimodal signals on hyper-complex space for extreme abstractive text summarization (tl;dr) of scientific contents. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3724–3736, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Erion Çano and Ondřej Bojar. 2019. Keyphrase generation: A text summarization struggle. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 666–672, Minneapolis, Minnesota. Association for Computational Linguistics.

Wang Chen, Yifan Gao, Jiani Zhang, Irwin King, and Michael R. Lyu. 2019. Title-guided encoding for keyphrase generation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. LipKey: A large-scale news dataset for absent keyphrases generation and abstractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3427–3437, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Lee, Rakesh M. Verma, Avisha Das, and Arjun Mukherjee. 2020. Experiments in extractive summarization: Integer linear programming, term/sentence scoring, and title-driven models. *CoRR*, abs/2008.00140.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding large language models via directional stimulus prompting. *Preprint*, arXiv:2302.11520.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022. CiteSum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10922–10935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Vivek Srivastava, Savita Bhat, and Niranjan Pedanekar. 2023. A few good sentences: Content selection for abstractive text summarization. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 124–141, Cham. Springer Nature Switzerland.

Gregor Stiglic, Leon Kopitar, Lucija Gosak, Primoz Kocbek, Zhe He, Prithwish Chakraborty, Pablo Meyer, and Jiang Bian. 2023. Improving primary healthcare workflow using extreme summarization of scientific literature based on generative ai. *Preprint*, arXiv:2307.15715.

Hong Sun, Xue Li, Yinchuan Xu, Youkow Homma, Qi Cao, Min Wu, Jian Jiao, and Denis Charles. 2023. Autohint: Automatic prompt optimization with hint generation. *Preprint*, arXiv:2307.07415.

Shahbaz Syed, Khalid Al Khatib, and Martin Potthast. 2024. TL;DR progress: Multi-faceted literature exploration in text summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 195–206, St. Julians, Malta. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. X-scitldr: cross-lingual extreme summarization of scholarly documents. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22. ACM.

Sotaro Takeshita, Simone Paolo Ponzetto, and Kai Eckert. 2024. Rouge-k: Do your summaries have keywords? *Preprint*, arXiv:2403.05186.

Weng Tam, Xiao Liu, Kaixuan Ji, Lilong Xue, Jiahua Liu, Tao Li, Yuxiao Dong, and Jie Tang. 2023. Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13117–13130, Singapore. Association for Computational Linguistics.

Peggy Tang, Kun Hu, Lei Zhang, Junbin Gao, Jiebo Luo, and Zhiyong Wang. 2023. Topiccat: Unsupervised topic-guided co-attention transformer for extreme multimodal summarisation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6643–6652, New York, NY, USA. Association for Computing Machinery.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. *Preprint*, arXiv:2211.01910.

## A Case Study

To further illustrate the effectiveness of our proposed EAP method, we present a case study comparing the generated TLDRs from EAP and the baseline SP method in zero-shot setting using GPT-3.5.

Table 5 showcases an example from the SciTLDR dataset, where the source paper is about combining differentiable decision trees (DDTs) with supervised variational autoencoders (SVAEs) to enhance the interpretability of classification tasks.

The TLDR generated by the SP method fails to capture the core idea of the paper, instead focusing on individual components such as DDTs and SVAEs without highlighting their synergistic combination. In contrast, the EAP method, guided by the attention signals derived from the title and keywords, successfully identifies the key aspects of the work, even in the absence of demonstration examples. The generated TLDR concisely conveys the main contribution: the integration of DDTs and SVAEs to create an interpretable and competitive classifier+VAE for high-dimensional data.

Moreover, the EAP-generated TLDR achieves a significantly higher Rouge-1 score (46.51) compared to the SP-generated TLDR (27.03), indicating its superior quality and relevance to the gold TLDRs. This case study demonstrates how the EAP method effectively guides the LLM to focus on the most critical information, resulting in more accurate and informative summaries.

## B Prompts

The prompts employed in this work for the standard prompting, explicit guidance prompting, and external attention prompting methods in the zero-shot and few-shot settings are illustrated in Figures 4, 5, 6, 7 and Figures 8, 9, 10, 11, respectively.

| Title | Interpretable Classification via Supervised Variational Autoencoders and Differentiable Decision Trees |
|---|---|
| Source abstract | As deep learning-based classifiers are increasingly adopted in real-world applications, the importance of understanding how a particular label is chosen grows. Single decision trees are an example of a simple, interpretable classifier, but are unsuitable for use with complex, high-dimensional data. On the other hand, the variational autoencoder (VAE) is designed to learn a factored, low-dimensional representation of data, but typically encodes high-likelihood data in an intrinsically non-separable way. We introduce the differentiable decision tree (DDT) as a modular component of deep networks and a simple, differentiable loss function that allows for end-to-end optimization of a deep network to compress high-dimensional data for classification by a single decision tree. We also explore the power of labeled data in a supervised VAE (SVAE) with a Gaussian mixture prior, which leverages label information to produce a high-quality generative model with improved bounds on log-likelihood. We combine the SVAE with the DDT to get our classifier+VAE (C+VAE), which is competitive in both classification error and log-likelihood, despite optimizing both simultaneously and using a very simple encoder/decoder architecture. |
| TLDR-AU | We combine differentiable decision trees with supervised variational autoencoders to enhance interpretability of classification. |
| TLDR-PR1 | This paper proposes a hybrid model of a variational autoencoder composed with a differentiable decision tree, and an accompanying training scheme, with experiments demonstrating tree classification performance, neg. log likelihood performance, and latent space interpretability. |
| TLDR-PR2 | The paper tries to build an interpretable and accurate classifier via stacking a supervised VAE and a differentiable decision tree |
| Standard Prompting (SP) | |
| Generated TLDR | The paper introduces the differentiable decision tree (DDT) as a modular component of deep networks and a simple, differentiable loss function that allows for end-to-end optimization of a deep network to compress high-dimensional data for classification by a single decision tree, and explores the power of labeled data in a supervised VAE (SVAE) with a Gaussian mixture prior, which leverages label information to produce a high-quality generative model with improved bounds on log-likelihood. |
| Rouge-1 | 27.03 |
| External Attention Prompting (EAP) | |
| Attention Signals | **Strong:** the differentiable decision tree<br>**Medium:** a simple, interpretable classifier; the variational autoencoder; single decision trees; a supervised VAE; classification error<br>**Weak:** deep learning-based classifiers; high-dimensional data; high-likelihood data; labeled data |
| Generated TLDR | The differentiable decision tree (DDT) and supervised VAE (SVAE) are combined to create an interpretable and competitive classifier+VAE for high-dimensional data with improved bounds on log-likelihood. |
| Rouge-1 | 46.51 |

Table 5: A case study comparing the TLDRs generated by GPT-3.5 using the standard prompting method and external attention prompting method in the zero-shot setting. TLDR-AU refers to the gold TLDR written by author and TLDR-PR1 and TLDR-PR2 are gold TLDRs derived from the perspective of peer reviewers, according to the representation in SciTLDR. The attention signals that are mentioned in EAP generated TLDR are highlighted in red. The EAP method, guided by attention signals derived from the title and keywords, generates a more accurate and informative TLDR that captures the core idea of the paper, even without demonstration examples.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 4: The prompt for standard prompting in the zero-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the title and abstract of a scientific paper, as well as a list of keywords, generate a short one-sentence summary
that captures the key aspects of the paper. You should pay attention to the given title and keywords when summarizing.

## Input

### Title
[source title]

### Abstract
source abstract

### Keywords
[source keywords]

### One-sentence summary
```

Figure 5: The prompt for explicit guidance prompting in the zero-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper.
You should pay attention to the emphasized content in the input abstract when summarizing.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 6: The prompt for random attention prompting in the zero-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper. You should pay attention to the emphasized content in the input abstract when summarizing.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 7: The prompt for external attention prompting in the zero-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper.

## EXAMPLE 1

### Abstract
Due to the success of deep learning to solving a variety of challenging machine learning tasks, there is a rising interest in understanding loss functions for training neural networks from a theoretical aspect. Particularly, the properties of critical points and the landscape around them are of importance to determine the convergence performance of optimization algorithms. In this paper, we provide a necessary and sufficient characterization of the analytical forms for the critical points (as well as global minimizers) of the square loss functions for linear neural networks. We show that the analytical forms of the critical points characterize the values of the corresponding loss functions as well as the necessary and sufficient conditions to achieve global minimum. Furthermore, we exploit the analytical forms of the critical points to characterize the landscape properties for the loss functions of linear neural networks and shallow ReLU networks. One particular conclusion is that: While the loss function of linear networks has no spurious local minimum, the loss function of one-hidden-layer nonlinear networks with ReLU activation function does have local minimum that is not global minimum.

### One-sentence summary
We provide necessary and sufficient analytical forms for the critical points of the square loss functions for various neural networks, and exploit the analytical forms to characterize the landscape properties for the loss functions of these neural networks.

## EXAMPLE 2

### Abstract
Reinforcement learning in an actor-critic setting relies on accurate value estimates of the critic. However, the combination of function approximation, temporal difference (TD) learning and off-policy training can lead to an overestimating value function. A solution is to use Clipped Double Q-learning (CDQ), which is used in the TD3 algorithm and computes the minimum of two critics in the TD-target. We show that CDQ induces an underestimation bias and propose a new algorithm that accounts for this by using a weighted average of the target from CDQ and the target coming from a single critic. The weighting parameter is adjusted during training such that the value estimates match the actual discounted return on the most recent episodes and by that it balances over- and underestimation. Empirically, we obtain more accurate value estimates and demonstrate state of the art results on several OpenAI gym tasks.

### One-sentence summary
A method for more accurate critic estimates in reinforcement learning.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 8: The prompt for standard prompting in the few-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper. You should pay attention to the given title and keywords when summarizing.

## EXAMPLE 1

### Title
Critical Points of Linear Neural Networks: Analytical Forms and Landscape Properties

### Abstract
Due to the success of deep learning to solving a variety of challenging machine learning tasks, there is a rising interest in understanding loss functions for training neural networks from a theoretical aspect. Particularly, the properties of critical points and the landscape around them are of importance to determine the convergence performance of optimization algorithms. In this paper, we provide a necessary and sufficient characterization of the analytical forms for the critical points (as well as global minimizers) of the square loss functions for linear neural networks. We show that the analytical forms of the critical points characterize the values of the corresponding loss functions as well as the necessary and sufficient conditions to achieve global minimum. Furthermore, we exploit the analytical forms of the critical points to characterize the landscape properties for the loss functions of linear neural networks and shallow ReLU networks. One particular conclusion is that: While the loss function of linear networks has no spurious local minimum, the loss function of one-hidden-layer nonlinear networks with ReLU activation function does have local minimum that is not global minimum.

### Keywords
ReLU activation function; linear neural networks; critical points; loss functions

### One-sentence summary
We provide necessary and sufficient analytical forms for the critical points of the square loss functions for various neural networks, and exploit the analytical forms to characterize the landscape properties for the loss functions of these neural networks.

## EXAMPLE 2

### Title
Dynamically Balanced Value Estimates for Actor-Critic Methods

### Abstract
Reinforcement learning in an actor-critic setting relies on accurate value estimates of the critic. However, the combination of function approximation, temporal difference (TD) learning and off-policy training can lead to an overestimating value function. A solution is to use Clipped Double Q-learning (CDQ), which is used in the TD3 algorithm and computes the minimum of two critics in the TD-target. We show that CDQ induces an underestimation bias and propose a new algorithm that accounts for this by using a weighted average of the target from CDQ and the target coming from a single critic. The weighting parameter is adjusted during training such that the value estimates match the actual discounted return on the most recent episodes and by that it balances over- and underestimation. Empirically, we obtain more accurate value estimates and demonstrate state of the art results on several OpenAI gym tasks.

### Keywords
more accurate value estimates; Reinforcement learning; off-policy training; several OpenAI gym

### One-sentence summary
A method for more accurate critic estimates in reinforcement learning.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 9: The prompt for explicit guidance prompting in the few-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper. You should pay attention to the emphasized content in the input abstract when summarizing.

## EXAMPLE 1

### Abstract
Due to the success of deep ***learning to solving*** a variety of challenging machine learning tasks, there is a rising interest in understanding ***loss function***s for training neural networks from a theoretical aspect. Particularly, the properties of critical points and the landscape around them are of importance to determine the convergence performance of optimization algorithms. In this paper, we provide a necessary and sufficient characterization of the analytical forms for the critical points (as well as global minimizers) of the square ***loss function***s for *linear neural networks*. We show that the analytical forms of the critical points characterize the values of the corresponding ***loss function***s as well as the necessary and sufficient conditions to achieve global minimum. Furthermore, we exploit the analytical forms of the critical points to characterize the landscape properties for the ***loss function***s of *linear neural networks* and **shallow ReLU networks**. One particular conclusion is that: While the ***loss function*** of linear networks has no spurious local minimum, the ***loss function*** of **one-hidden-layer nonlinear networks** with ReLU **activation function** does have local minimum that is not global minimum.

### One-sentence summary
We provide necessary and sufficient analytical forms for the critical points of the square loss functions for various neural networks, and exploit the analytical forms to characterize the landscape properties for the loss functions of these neural networks.

## EXAMPLE 2

### Abstract
Reinforcement learning in an actor-critic setting relies on accurate value estimates of the critic. However, the combination of *function approximation*, temporal difference (TD) learning and off-policy **training can lead** to an overestimating value function. A solution is to use Clipped Double Q-learning (CDQ), which is used in the TD3 *algorithm and computes* the minimum of two critics in the TD-target. We show that CDQ induces an underestimation bias and propose a new algorithm that accounts for this by ***using a weighted average*** of the target from CDQ and the target coming from a single critic. The weighting parameter is adjusted during training such that the value estimates match the actual discounted return on the most recent episodes and by that it balances over- and underestimation. Empirically, we obtain more accurate value estimates and demonstrate state of the art results on several ***OpenAI gym*** tasks.

### One-sentence summary
A method for more accurate critic estimates in reinforcement learning.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 10: The prompt for random attention prompting in the few-shot setting.

```
# Task: Extreme summarization of scientific documents

## Task Definition
Given the abstract of a scientific paper, generate a short one-sentence summary that captures the key aspects of the paper. You should pay attention to the emphasized content in the input abstract when summarizing.

## EXAMPLE 1

### Abstract
Due to the success of deep learning to solving a variety of challenging machine learning tasks, there is a rising interest in understanding *loss functions* for training neural networks from a theoretical aspect. Particularly, the properties of ***critical points*** and the landscape around them are of importance to determine the convergence performance of optimization algorithms. In this paper, we provide a necessary and sufficient characterization of **the analytical forms** for the ***critical points*** (as well as global minimizers) of the square *loss functions* for ***linear neural networks***. We show that **the analytical forms** of the ***critical points*** characterize the values of the corresponding *loss functions* as well as the necessary and sufficient conditions to achieve global minimum. Furthermore, we exploit **the analytical forms** of the ***critical points*** to characterize **the landscape properties** for the *loss functions* of ***linear neural networks*** and shallow ReLU networks. One particular conclusion is that: While the loss function of linear networks has no spurious local minimum, the loss function of one-hidden-layer nonlinear networks with *ReLU activation function* does have local minimum that is not global minimum.

### One-sentence summary
We provide necessary and sufficient analytical forms for the critical points of the square loss functions for various neural networks, and exploit the analytical forms to characterize the landscape properties for the loss functions of these neural networks.

## EXAMPLE 2

### Abstract
*Reinforcement learning* in **an actor-critic setting** relies on accurate ***value estimates*** of the critic. However, the combination of function approximation, temporal difference (TD) learning and *off-policy training* can lead to an overestimating value function. A solution is to use Clipped Double Q-learning (CDQ), which is used in the TD3 algorithm and computes the minimum of two critics in the TD-target. We show that CDQ induces an underestimation bias and propose a new algorithm that accounts for this by using a weighted average of the target from CDQ and the target coming from a single critic. The weighting parameter is adjusted during training such that the ***value estimates*** match the actual discounted return on the most recent episodes and by that it balances over- and underestimation. Empirically, we obtain ***more accurate value estimates*** and demonstrate state of the art results on *several OpenAI gym* tasks.

### One-sentence summary
A method for more accurate critic estimates in reinforcement learning.

## Input

### Abstract
[source abstract]

### One-sentence summary
```

Figure 11: The prompt for external attention prompting in the few-shot setting.