

# News Recommendation with Category Description by a Large Language Model

Yuki Yada<sup>1\*</sup>, Hayato Yamana<sup>1</sup>

<sup>1</sup>Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku Tokyo 169-8050 Japan

## Abstract

Personalized news recommendations are essential for online news platforms to assist users in discovering news articles that match their interests from a vast amount of online content. Appropriately encoded content features, such as text, categories, and images, are essential for recommendations. Among these features, news categories, such as *tv-golden-globe*, *finance-real-estate*, and *news-politics*, play an important role in understanding news content, inspiring us to enhance the categories' descriptions. In this paper, we propose a novel method that automatically generates informative category descriptions using a large language model (LLM) without manual effort or domain-specific knowledge and incorporates them into recommendation models as additional information. In our comprehensive experimental evaluations using the MIND dataset, our method successfully achieved 5.8% improvement at most in AUC compared with baseline approaches without the LLM's generated category descriptions for the state-of-the-art content-based recommendation models including NAML, NRMS, and NPA. These results validate the effectiveness of our approach. The code is available at <https://github.com/yamanalab/gpt-augmented-news-recommendation>.

## Keywords

News Recommendation, Large Language Model, Deep Neural Network

## 1. Introduction

In recent years, online news platforms have gained widespread popularity, making it commonplace for users to consume news articles on their mobile devices. These platforms deliver a large amount of news content to users daily, leading to an information overload problem, making it challenging to discover articles that align with user's interests. Therefore, personalized news recommendations have become essential for online news platforms.

Many proposed news recommendation models use deep neural networks, which achieve high performance [1, 2, 3, 4, 5, 6]. As shown in Figure 1, most neural news recommendation models adopt a common approach that consists of three core modules:

1. News encoder: it generates a news vector that captures the semantic information from the news content.
2. User encoder: it generates a user vector that captures the user preferences based on browsing history.
3. Similarity calculator: it computes the similarity score between the news vector and the user vector, estimating that news items with higher similarity scores are more likely to match the user's interests.

Previous studies have focused on investigating which neural network structures to apply in the user and news encoder modules, resulting in various proposed architectures for news recommendation models. Recent methods incorporate pre-trained language models (PLMs) into the news encoder to learn news representations. These PLM-based approaches have achieved high performance [7, 2, 8, 6, 9].

Simultaneously, a crucial source for understanding news content is the news category. For example, the MIND [10] dataset, a well-known dataset in the news recommendation field, contains news categories

---

INRA 2025: 12th International Workshop on News Recommendation and Analytics in Conjunction with ACM RecSys 2025

✉ [yada.yuki@fuji.waseda.jp](mailto:yada.yuki@fuji.waseda.jp) (Y. Yada); [yamana@yama.info.waseda.ac.jp](mailto:yamana@yama.info.waseda.ac.jp) (H. Yamana)

🌐 <https://www.yyada.jp/> (Y. Yada)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

such as *tv-golden-globes*. The naïve way to use categories in recommendation models is to adopt predefined templates (e.g., *The news category is {category}*) followed by combining them with original news recommendation models. However, these templates are too generic to be applied to all news categories, resulting in insufficient information. Constructing detailed descriptions for each news category and using them as input can be beneficial for enabling recommendation models to recognize news content accurately.

However, manually constructing detailed descriptions for each news category is costly. To address this issue, we propose automatically generating descriptive text for news categories using large language models (LLMs). Since LLMs are pre-trained on vast amounts of text data and have extensive knowledge across diverse topics, the LLM-generated detailed descriptions are expected to enhance the recommendation performance. The contributions of this paper are as follows:

1. We first propose the adoption of LLM-generated news category descriptions to enhance news recommendations.
2. Comprehensive experimental evaluations on the MIND dataset confirmed that our proposed method consistently outperforms baseline approaches across multiple recommendation models, achieving up to 5.8% improvement in AUC compared to methods without category descriptions.

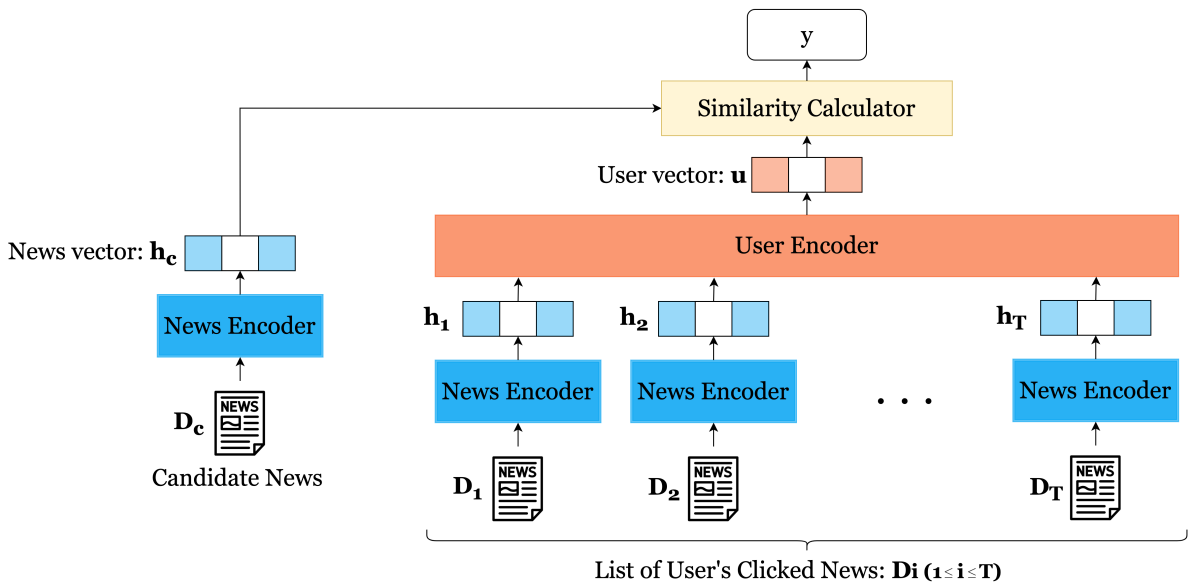
The rest of this paper is organized as follows. In Section 2, we reviews state-of-the-art news recommendations and use of LLM-generated text. We propose a new recommendation method enhanced by an LLM in Section 3, followed by evaluations in Section 4. In Section 5, we discuss the limitations of the proposed method. Finally, In Section 6, we conclude this paper.

## 2. Related Work

In this section, we review related work on news recommendations and the use of LLM-generated text.

### 2.1. News Recommendations

News recommendations have been extensively studied through various models [2, 6]. Neural network-based models have achieved state-of-the-art performance by encoding news content and user preferences to capture the complex interactions between users and news items [1, 11, 3, 4, 5]. For instance, NRMS



**Figure 1:** Neural News Recommendation Model

[5] employs multi-head attention to acquire news and user vectors, while LSTUR [1] adopts GRU for to obtain user vectors. Other models, such as MCCM [11], NAML [3], and NPA [4], leverage CNN and attention mechanisms to generate user and news vectors.

Particularly, the models adopting PLMs, such as BERT [12] and RoBERTa [13], to obtain news vectors have achieved notably high performance [7, 2, 8, 6, 9, 14]. For instance, Wu et al. [9] proposed a framework incorporating PLMs into news recommendations, demonstrating 2.63% improvement in the pageview rate through online experiments. In the ONCE [7], proposed by Liu et al., showed that using large-scale language models such as LLaMa [15] can improve news recommendation performance.

Several neural recommendation models use different architectures from the model shown in Figure 1. For instance, Zhang et al. introduced UNBERT [16], which feeds a candidate news article and a list of previously viewed news articles into a single language model to predict click-through rates. Another notable work by Zhang et al. [17] is Prompt4NR, a news recommendation approach using prompt learning with PLMs.

In summary, various methods with neural network-based models have been proposed for news recommendation. The use of PLMs shows particularly high performance; however, PLMs lack sufficient knowledge to interpret category names. Thus, a space exists to enhance the performance of news recommendations by adding detailed category information, which inspires us to adopt LLM-generated category descriptions as additional features.

## 2.2. Use of LLM-Generated Texts

The use of text generated by LLMs, which have acquired extensive knowledge on a wide range of topics through pre-training, has improved the performance in various natural language processing tasks, such as text classification and question answering [18, 19, 20].

Yoo et al. [20] proposed GPT3Mix, which applies GPT-3-based [21] data augmentation to text classification tasks by generating additional training examples. For question answering tasks, Liu et al. [18] used knowledge expansion using GPT-3 to augment the context and improve the performance of the question answering model. Pratt et al. [19] introduced CuPL, a method that leverages image captions generated by LLMs for zero-shot image classification using CLIP [22]. They automatically generated image captions using LLMs and used them as text inputs for CLIP.

These studies demonstrated the effectiveness of utilizing texts generated by LLMs in various tasks. Inspired by this, we aim to improve news recommendation performance by generating news category descriptions using LLMs.

## 3. Proposed Method

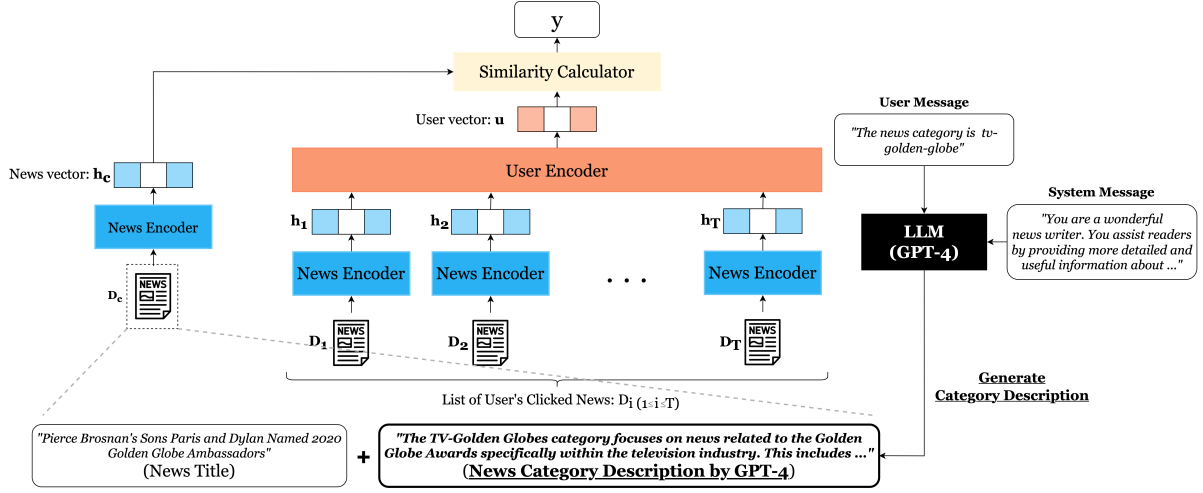
Figure 2 shows our proposed novel news recommendation method enhanced by LLM-generated category descriptions, consisting of two steps: 1) automatic generation of news category descriptions, and 2) integration of the generated category descriptions with the recommendation model as additional input features.

### 3.1. Generation of Category Descriptions

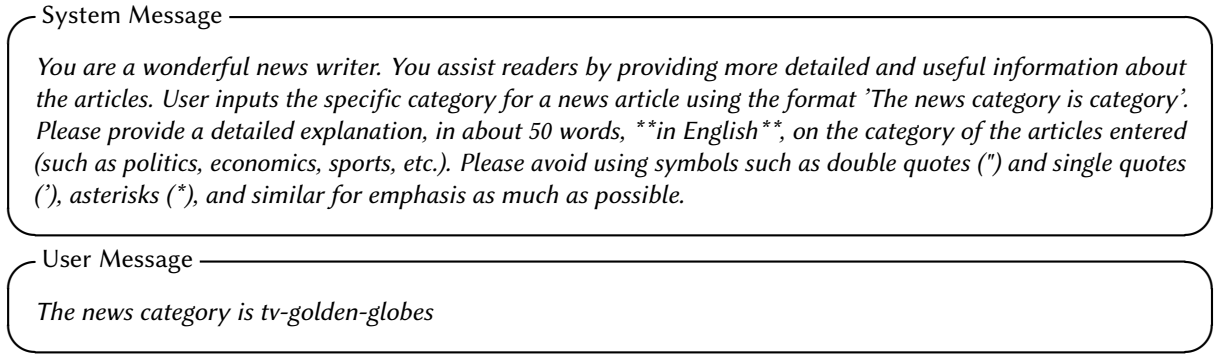
In this step, we generate the descriptions for news categories using an LLM without any manual effort. We employ GPT-4 [23] as the LLM for generating category descriptions, as it has demonstrated high performance across various domains and tasks [24].

Figure 3 shows the prompts for the LLM to generate descriptions for news categories. Although the prompts need to be prepared manually, they do not require any manual effort once they have been prepared.

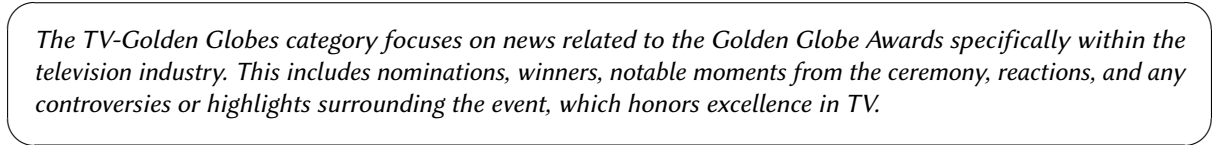
Figure 4 shows a specific example of the generated news category descriptions for the *tv-golden-globes* category, one of the categories that appeared in the MIND dataset [10]. The average word count of the



**Figure 2:** Overview of our proposed method



**Figure 3:** Prompt to generate a category description for *tv-golden-globe* as an example.



**Figure 4:** Category description for *tv-golden-globes* via GPT-4

generated category descriptions is 57.7 for all 270 news categories within the MIND dataset used in our experiment.

### 3.2. Integration of Category Descriptions into News Recommendation Models

After generating the category descriptions, we train the news recommendation model by leveraging the generated descriptions. Let  $D_{title}$  denote the news title text and  $D_{desc}$  represent the generated category description. During the inference phase of the recommendation model, we concatenate  $D_{title}$  and  $D_{desc}$  using the special *SEP* token from BERT [12], and feed the resulting text into the news encoder.

## 4. Experiments and Results

In this section, we describe our experiments. First, we present the dataset, evaluation metrics, baselines, and hyperparameters used in our experiments. Second, we present the results and then discuss the

**Table 1**

MIND dataset [10]

#User	#News	#Impression	#Click
94,057	65,238	230,117	347,727

**Table 2**

Comparison of different methods on the MIND dataset

Model	PLM	Method	AUC	MRR	nDCG@5	nDCG@10
NAML [3]	DistilBERT	<i>title only</i>	0.675	0.292	0.317	0.384
		<i>title + template-based</i>	0.690	0.295	0.327	0.393
		<i>title + generate-description (ours)</i>	<b>0.713</b>	<b>0.326</b>	<b>0.363</b>	<b>0.425</b>
	BERT	<i>title only</i>	0.700	0.318	0.350	0.414
		<i>title + template-based</i>	0.696	0.308	0.340	0.405
		<i>title + generate-description (ours)</i>	<b>0.707</b>	<b>0.322</b>	<b>0.357</b>	<b>0.420</b>
NRMS [5]	DistilBERT	<i>title only</i>	0.674	0.297	0.322	0.387
		<i>title + template-based</i>	0.675	0.311	0.341	0.400
		<i>title + generate-description (ours)</i>	<b>0.707</b>	<b>0.324</b>	<b>0.359</b>	<b>0.422</b>
	BERT	<i>title only</i>	0.689	0.306	0.336	0.400
		<i>title + template-based</i>	0.667	0.301	0.329	0.389
		<i>title + generate-description (ours)</i>	<b>0.706</b>	<b>0.320</b>	<b>0.355</b>	<b>0.418</b>
NPA [4]	DistilBERT	<i>title only</i>	0.700	0.311	0.344	0.408
		<i>title + template-based</i>	0.698	0.309	0.342	0.407
		<i>title + generate-description (ours)</i>	<b>0.707</b>	<b>0.319</b>	<b>0.354</b>	<b>0.417</b>
	BERT	<i>title only</i>	0.689	0.301	0.332	0.398
		<i>title + template-based</i>	0.694	0.314	0.345	0.410
		<i>title + generate-description (ours)</i>	<b>0.710</b>	<b>0.324</b>	<b>0.360</b>	<b>0.422</b>

effectiveness of the proposed method.

#### 4.1. Experimental Setting

We used the MIND [10] dataset, a widely used dataset for the news recommendation field. The MIND dataset, whose statistics are shown in Table 1, was constructed from the news click records of the Microsoft News<sup>1</sup> service between October 12, 2019 and November 22, 2019.

We compared our proposed method with the following two baselines:

1. *title only*: Input only the title text into the news encoder without any category information.
2. *title + template-based*: Input the concatenated text of the title and the template-based sentence that follows the template: *The news category is {category}* (e.g., *The news category is tv-golden-globes*), using the *SEP* token.

To evaluate the effectiveness of our proposed method, we adopted three state-of-the-art content-based recommendation models: NAML [3], NRMS [5], and NPA [4]. We used the pre-trained models DistilBERT-base (*distilbert-base-uncased*) [25] and BERT-base (*bert-base-uncased*) [12] for the news encoder component. AdamW [26] was used as the optimization function with a learning rate of 1e-4, a batch size of 128, and 3 epochs. Following the previous studies [11, 4, 9, 17], the maximum number of recently viewed articles, which was used to capture user preference information, was set to 50. We set the number of negative samples in the training phase to 4. All experiments are conducted on a Tesla V100 GPU. For implementation, we used PyTorch v2.1.0 [27] and Transformers v4.35.0 [28] libraries. As performance metrics, we used the average AUC, MRR, nDCG@5, and nDCG@10 for all impressions.

<sup>1</sup><https://news.microsoft.com>

**Table 3**News titles for the *tunedin* category

News title
<i>Coca-Cola released two limited-edition holiday flavored sodas</i>
<i>Google Maps and Waze may share certain features</i>
<i>Daylight saving time ends: Get ready to 'fall back'</i>

*The category "tunedin" typically refers to news related to entertainment, media, and television. It often includes updates on TV shows, interviews with celebrities, insights into the music industry, and information on streaming services. This category keeps readers engaged with the latest trends and happenings in the world of entertainment.*

**Figure 5:** Category description for the *tunedin* category

## 4.2. Results

Table 2 summarizes the comparison of our proposed recommendation method with the baselines. The proposed method, *title+generated-description*, demonstrates the highest performance compared with the baselines across all recommendation models and achieved up to 5.6% and 5.8% improvements in AUC compared with the *title only* and *title+template-based* baselines, respectively.

Furthermore, *title + template-based* shows insignificant improvement in performance compared with the *title only* in most cases. This result suggests that simply fitting the category name into a predefined template is ineffective for representing the category information. The results also indicate that for relatively small-scale language models like BERT, using only the category names as input does not provide sufficient information for the recommendation models to properly interpret the category information.

## 5. Limitations

To explore the limitations, we performed a manual inspection of the category descriptions. Although GPT-4 generally produces high-quality descriptions, it occasionally fails to generate accurate descriptions.

Figure 5 shows an example in which the generated description for the *tunedin* category focuses on entertainment, music, and television. However, the actual article titles in Table 3 reflect a wider range of topics, including technology and trends. A more accurate description would be that the *tunedin* category casually provides news on various topics such as entertainment, technology, and business, keeping readers updated on the latest trends. This discrepancy highlights the model’s inability to accurately capture the category’s broad scope.

This example, along with other similar instances, suggests that the model may struggle to generate accurate category descriptions when the input lacks sufficient background knowledge or context.

## 6. Conclusion

In this study, we proposed a method to automatically generate news category descriptions using LLMs and incorporate them into news recommendation models. The experiments on the MIND dataset demonstrated that our method outperformed the baselines across all metrics for multiple models. Our main contribution is a novel approach to enhance news recommendation models’ understanding of category information using LLMs. Our future work will include enhancing the recommendation performance by improving the generated descriptions.

## Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini in order to assist with translation and to review and improve the clarity of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

- [1] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, X. Xie, Neural News Recommendation with Long- and Short-term User Representations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 336–345.
- [2] X. Meng, H. Huo, X. Zhang, W. Wang, J. Zhu, A Survey of Personalized News Recommendation, *Data Science and Engineering* (2023) 1–21.
- [3] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, X. Xie, Neural news recommendation with attentive multi-view learning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, Macao, China, 2019, p. 3863–3869.
- [4] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, X. Xie, NPA: Neural News Recommendation with Personalized Attention, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 2019, pp. 2576–2584.
- [5] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, X. Xie, Neural News Recommendation with Multi-Head Self-Attention, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 6389–6394.
- [6] C. Wu, F. Wu, Y. Huang, X. Xie, Personalized News Recommendation: Methods and Challenges, *ACM Trans. Inf. Syst.* 41 (2023) 1–50.
- [7] Q. Liu, N. Chen, T. Sakai, X.-M. Wu, ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models, *arXiv preprint arXiv:2305.06566* (2023).
- [8] K. Shivaram, P. Liu, M. Shapiro, M. Bilgic, A. Culotta, Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation, in: Proceedings of the 16th ACM Conference on Recommender Systems, Seattle, WA, USA, 2022, pp. 220–228.
- [9] C. Wu, F. Wu, T. Qi, Y. Huang, Empowering News Recommendation with Pre-Trained Language Models, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 2021, pp. 1652–1656.
- [10] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, M. Zhou, MIND: A Large-scale Dataset for News Recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 2020, pp. 3597–3606.
- [11] J. Wang, Y. Jiang, H. Li, W. Zhao, Improving News Recommendation with Channel-Wise Dynamic Representations and Contrastive User Modeling, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, Singapore, 2023, pp. 562–570.
- [12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019, pp. 4171–4186.
- [13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv:1907.11692* (2019).
- [14] C. Wu, F. Wu, T. Qi, C. Zhang, Y. Huang, T. Xu, MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022, pp. 2560–2564.
- [15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal,



- E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, arXiv preprint arXiv:2302.13971 (2023).
- [16] Q. Zhang, J. Li, Q. Jia, C. Wang, J. Zhu, Z. Wang, X. He, UNBERT: User-News Matching BERT for News Recommendation, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, Virtual, 2021, pp. 3356–3362.
  - [17] Z. Zhang, B. Wang, Prompt Learning for News Recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 2023, pp. 227–237.
  - [18] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, H. Hajishirzi, Generated knowledge prompting for commonsense reasoning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022, pp. 3154–3169.
  - [19] S. Pratt, I. Covert, R. Liu, A. Farhadi, What does a platypus look like? Generating customized prompts for zero-shot image classification, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 15645–15655.
  - [20] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, W. Park, GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 2021, pp. 2225–2239.
  - [21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Vancouver, BC, Canada, 2020, pp. 1877–1901.
  - [22] R. Alec, W.-K. Jong, H. Chris, R. Aditya, G. Gabriel, A. Sandhini, S. Girish, A. Amanda, M. Pamela, C. Jack, K. Gretchen, S. Ilya, Learning Transferable Visual Models From Natural Language Supervision, arXiv preprint arXiv:1907.11692 (2021).
  - [23] OpenAI, GPT-4 Technical Report, arXiv preprint arXiv:2303.08774 (2023) (2023).
  - [24] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of Artificial General Intelligence: Early experiments with GPT-4, arXiv preprint arXiv:2303.12712 (2023).
  - [25] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2020).
  - [26] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
  - [27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019, pp. 8024–8035.
  - [28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Virtual, 2020, pp. 38–45.