Enhancing Personalized Headline Generation via Offline Goal-conditioned Reinforcement Learning with Large Language Models

Xiaoyu Tan*
txywilliam1993@outlook.com
INF Technology (Shanghai) Co., Ltd.
Shanghai, China

Leijun Cheng*
leijuncheng@sues.edu.cn
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science
Shanghai, China

Xihe Qiu*†
qiuxihe1993@gmail.com
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science
Shanghai, China

Shaojie Shi
tjpolyurethane@gmail.com
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science
Shanghai, China

Yuan Cheng
cheng_yuan@fudan.edu.cn
AI^3 Institute, Fudan University
Shanghai, China

Wei Chu
chuwei@inftech.ai
INF Technology (Shanghai) Co., Ltd.
Shanghai, China

Yinghui Xu
xuyinghui@fudan.edu.cn
AI^3 Institute, Fudan University
Shanghai, China

Yuan Qi
qi.yuan@outlook.com
AI^3 Institute, Fudan University
Shanghai, China

ABSTRACT

Recently, significant advancements have been made in Large Language Models (LLMs) through the implementation of various alignment techniques. These techniques enable LLMs to generate highly tailored content in response to diverse user instructions. Consequently, LLMs have the potential to serve as robust, customizable recommendation systems in the field of content recommendation. However, using LLMs with user individual information and online exploration remains a challenge, which are important perspectives in developing personalized news headline generation algorithms. In this paper, we propose a novel framework to generate personalized news headlines using LLMs with extensive online exploration. The proposed approach involves initially training an offline goal-conditioned policy using supervised learning. Subsequently, online exploration is employed to collect new data for the next training iteration. Results from simulations, experiments, and real-word scenario demonstrate that our framework achieves outstanding performance on established benchmarks and can effectively generate personalized headlines under different reward settings. By treating the LLM as a goal-conditioned agent, the model can perform online exploration by modifying the goals without frequently retraining the model. To the best of our knowledge, this work represents the first investigation into the capability of LLMs to generate customized news headlines with goal-conditioned reinforcement learning via supervised learning within LLMs.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; **Natural language processing**.

KEYWORDS

News Headline Generation, Personality, Large Language Models

ACM Reference Format:
Xiaoyu Tan, Leijun Cheng, Xihe Qiu, Shaojie Shi, Yuan Cheng, Wei Chu, Yinghui Xu, and Yuan Qi. 2024. Enhancing Personalized Headline Generation via Offline Goal-conditioned Reinforcement Learning with Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671638

1 INTRODUCTION

In recent times, we have observed a significant advancement in the development of large language models (LLMs) through the utilization of various alignment techniques [8, 9, 13, 49, 51]. These techniques aim to align the LLMs as conversational assistants, such as ChatGPT and Claude [32], capable of comprehending diverse instructions from humans. Typically, these models undergo a two-phase learning process. Firstly, they are trained through supervised

*Authors contributed equally to this research.
†Corresponding author

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671638

fine-tuning (SFT) [15, 23, 36] using a curated set of instructions to learn how to follow them accurately. Subsequently, reinforcement learning from human feedback (RLHF) [12, 33] is employed to further refine the models and incorporate human values. This learning procedure enables the LLMs to attain human-expert level across multiple domains in content generation.

However, current research is primarily focused on acquiring a diverse and comprehensive skill set for LLMs [47], with less emphasis on exploring alignment techniques within personalized and customized industrial domains. A well-established industrial application in this context is personalized news headline generation, which involves generating news headlines tailored to user interests, news content, user preferences, and viewing histories. Additionally, news headline generation entails economic and ethical constraints, as well as the need for fast adaptation to current events and changes in user preferences. Therefore, the direct application of existing alignment techniques cannot adequately address the challenges specific to the personalized news headline generation domain.

In this paper, we propose a novel learning framework to enhance personalized news headline generation with **o**ffline **g**oal-conditioned reinforcement learning (RL) and **o**nline **e**xploration (OGOE) using LLMs. Initially, we carefully structure user information and behaviors into a natural language prompt, which is then used to fine-tune the LLM using SFT. This process enables the LLM to follow the initial instructions accurately. Next, we employ an offline goal-conditioned RL method via supervised learning to incorporate diverse perspectives and feedback from reward signals. This approach eliminates the requirement for complex training processes in RLHF. Lastly, we design an online exploration technique that takes into account the economic cost by changing the goal, thus minimizing the requirement for frequent retraining of the LLM to get different policies.

We assess the efficacy of our proposed framework by conducting evaluations on publicly available benchmarks, custom-designed simulations, and one real-world financial report headline generation task. Specifically, we evaluate the ability of our framework to generate personalized headlines using goal-conditioned RL via supervised learning and online exploration. The experimental results showcase the superior performance and robust generalization capability of the LLM under our proposed framework. The contributions of our work are outlined as follows:

- We present a novel theoretical framework for accomplishing personalized news headline generation by employing offline goal-conditioned RL and online exploration in conjunction with LLMs.
- To the best of our knowledge, this study represents a pioneering effort in evaluating the offline goal-conditioned RL capabilities of LLMs through supervised learning.
- Our framework demonstrates exceptional performance in publicly available benchmarks , designed simulations, and real-world scenario.
- Our investigation reveals that the LLMs exhibit notable capacity in generalization as goal-conditioned agent via supervised learning, a phenomenon that has not been previously documented in existing literature.

2 RELATED WORKS

Headline generation is widely recognized as a specialized form of text summarization [28]. Both tasks involve methods for extracting[2, 14] and abstracting [18, 41, 44, 46, 48] information from the source text, and have gained wide recognition and extensive research in the field of natural language processing (NLP). However, distinct differences exist between them. The objective of text summarization is to summarize and extract information from the source text, providing a concise overview of the key points. In contrast, news headline generation not only entails summarizing information but also requires consideration of user interaction and preferences. The goal is to help different users quickly grasp the most relevant points and essential content of a news article.

The development of headline generation and text summarization techniques is significant for improving applications such as information retrieval, document understanding, and content summarization. Extraction-based methods that directly select sentences from the source text can lead to incoherent summaries [2]. On the other hand, abstractive models using an encoder-decoder framework [5, 6, 30] can generate more concise and fluent outputs based on the deeper semantics of the news content.

In recent years, approaches to generate personalized news headlines through implicit style transfer [6, 17, 35] or supervised outputs guided by a particular style have emerged. However, training personalized models for each user is infeasible due to the complexity of personalized linguistic styles and the scarcity of personalized examples. Meanwhile, training a unified text-style transfer model fails to meet the personalized needs of different users. Additionally, these methods face the risk of generating headlines purely for eye-catching purposes rather than delivering genuine value. Therefore, building a personalized text summarization or news summarization model that accounts for personalization remains a significant challenge.

While LLMs exhibit remarkable advantages in various aspects [37], the use of such models for generating personalized news headlines carries the risk of amplifying biases present in the training data, potentially resulting in biased or unfair news headlines [16]. Furthermore, the utilization of user data for personalized news headline generation raises concerns regarding user privacy and the careful handling of sensitive information [39]. Therefore, there is a need for a dedicated method to generate personalized news headlines.Distinct from prior methodologies, we aim to employ the news information previously clicked by users as personalized features, utilizing offline objective condition strategies and exploration to develop a personalized news headline generation model.This approach has the capability to better align with users' personalized information while ensuring the reality of generated headlines.

3 METHODS

In the OGOE framework, we first perform SFT on personalized news generation data to ensure the initial instruction-following capability of the model. Then, we perform reinforcement learning to optimize the quality of personalized generation. Different from the previous works that utilized on-policy reinforcement learning algorithms [4] to fine-tune the whole model, we perform goal-conditioned reinforcement learning solely through supervised learning to avoid

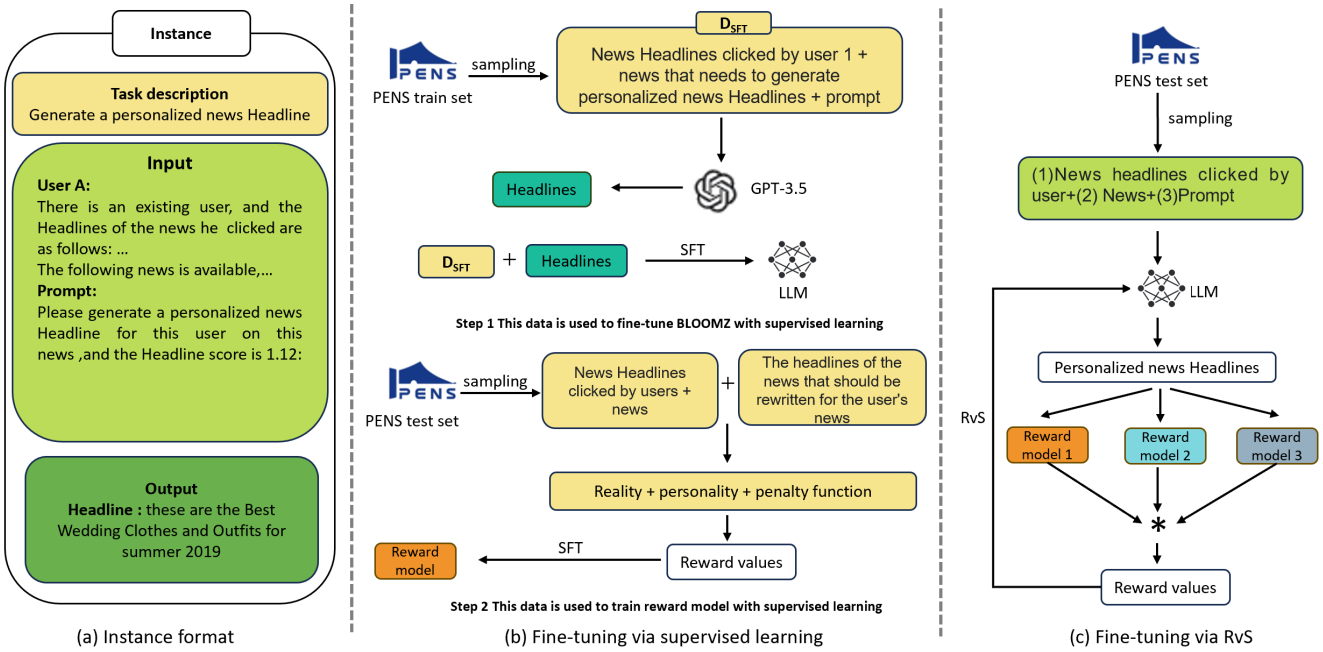

Figure 1: An overview of our proposed OGOE for personalized news headline generation.

complex model parallel optimization processes with improved performance. Finally, by in-cooperating online exploration strategies with budget limitations by only tuning the conditioned reward value, the model can perform effective personalized news headline generation, which efficiently avoids the need for frequent online fine-tuning.

3.1 Preliminary

In the personalized news headline generation tasks, typically, we can observe specific user historical behaviors $\mathbf{X} = [X_{user}, (X^0_{news}, Y^0_{headline})...(X^n_{news}, Y^n_{headline})]$ on i-th behaviors and user personal features X_{user}. For a given news X^{n+1}_{news}, we should train a news headline generation model P that can generate personalized news headlines condition on all user information $Y^{n+1}_{headline} \sim P(\cdot|\mathbf{X})$.Over the past few years, methods have been developed to create news headlines by implicitly transferring style [6, 17, 35].

3.2 Supervised Fine-tuning on News Headline Generation Set

Since the LLMs are not aligned to perform the aforementioned task, we utilize a dataset D_{SFT} to organize the data point by aggregating the user's personal features and news viewing histories. For each viewing history, we set the current news title as $Y^{n+1}_{headline}$ with number of n viewing histories that the time stamp in log is smaller than the viewing behavior of current news. To formulate the input, we design an input prompt $prompt_{SFT}$ that organizes the user features and viewing histories in readable natural language. The prompt $prompt_{SFT}$ is shown in the Section 3.3. Then, in dataset D_{SFT}, we have data pair $(\mathbf{X}_{SFT} = prompt_{SFT}(\mathbf{X}, X^{n+1}_{news}), Y^{n+1}_{headline})$. Finally, we train the model P_θ with parameter θ by maximizing the log probability of each token in $Y^{n+1}_{headline}$ auto-repressively:

$$\max_\theta \mathbb{E}_{D_{SFT}} \log P_\theta(Y^{n+1}_{headline}|\mathbf{X}_{SFT}). \quad (1)$$

Typically, we can achieve this by minimizing the cross-entropy loss [53] with the Adam optimizer [25] and getting the model θ_{SFT}.

3.3 Prompt Design

To enable the generation of personalized news headlines tailored to individual user interests, we have designed a structured prompt format that conditions the language model with precision:

$prompt_{SFT}$

There is an existing user,and the Headlines of the news he clicked are as follows:
[user's click history containing N previously clicked news headlines]
The following news is available, [the source news article].
Please generate a personalized news headline for user on this news, and the Headline score is X:

Specifically, the prompt consists of three pivotal components:

(1) The user's click history on N news headlines, encapsulating their interests.
(2) The full text of the source news article, providing contextual information.

(3) A request to generate a personalized headline with personalization score X, explicitly specifying the desired level of customization.

Feeding this structured prompt conditions the model to produce customized headlines by accounting for both user preferences and article content. The personalization score offers tunable control over the degree of tailored generation.

3.4 Reinforcement Learning Via Supervised Learning

To align the LLM effectively with different desirable values, multiple reinforcement learning methods are developed to optimize the capability of instruction following. For example, RLHF performs on-policy optimization by independently training a reward model and using it as a simulator to fine-tune the model. By modeling the reward that is labeled by the human annotators, the reward models can somehow represent the human values of helpfulness, harmlessness, and honesty. However, performing on-policy RL training (e.g., the proximal policy gradient algorithm) requires parallel caching of the base policy model, reward model, and optimization model. Hence, it requires a large quantity of resources, especially for the online exploration and updating scenario, and is not suitable for news title generation tasks.

In the OGOE framework, we perform goal-conditioned RL to learn personalized news headline generation with different values. We leverage the Reinforcement Learning via Supervised Learning (RvS) framework to design the personalized value alignment process. Given the dataset D_{RL} that contains data $(\mathbf{X}, X^{n+1}_{news}, Y^{n+1}_{headline}, r^{n+1})$, where r^{n+1} is the reward value that is related to the specific scenario definition. We train the policy model $\pi_{\theta'}$ with parameter θ' by optimizing: where the π is initialized by θ_{SFT} and $prompt_{RL}$ is

$$\max_{\theta'} \mathbb{E}_{D_{RL}} \log \pi_{\theta'}(Y^{n+1}_{headline} | prompt_{RL}(\mathbf{X}, X^{n+1}_{news}, r^{n+1})), \quad (2)$$

the prompt. $prompt_{RL}$ is similar to the $prompt_{SFT}$ but with varied reward values r_{target}. For inference, we can set a desirable goal target r_{target} and sample the news headlines $Y_{headline} \sim \pi_{\theta'}(\cdot | prompt_{RL}(\mathbf{X}, X_{news}, r = r_{target}))$. Here, the use of goal target r_{target} is varied; for exploitation, for example, we can set the r_{target} as the 80-th percentile of the maximum value of rewards in D_{RL} to enhance the robustness. For online exploration and budget allocation, we can shape the r_{target} and the details would be introduced in the Section 3.5.

3.5 Exploration without Online Updates

Due to the fast online distributional shift of daily news contents, the personalized news headline generation model should be updated frequently with the capability of exploration. However, as a practical system used in industrial scenarios, the exploration process should also consider budget limitations. To incorporate the exploration through reward uncertainties [10] and the goal-conditioned reinforcement learning via supervised learning, we can effectively perform online exploration without frequently re-training the news headline generation model, which is built upon LLMs. The online budget can also be controlled through a goal-conditioned policy $\pi_{\theta'}$ by tuning the goal setting automatically.

To accurately approximate the reward uncertainty, we can independently train M reward models $\{r_{\Phi_i}\}_i^M$ that only perform reward value regression using equation (1) with $\hat{r}_i = r_{\Phi_i}(\mathbf{X}, X_{news}, Y_{headline})$. In practice, we can implement low-rank adaptation (LoRA) [20] to train r_{Φ_i} which is a parameter-efficient framework, upon the LLMs without directly fine-tuning them. Similar to [11], we can get the extrinsic reward r^{ext}, intrinsic reward r^{int}, and the total goal reward r^{total} by:

$$\begin{aligned} r^{ext} &= \frac{1}{M}\sum_{i=1}^{M} \hat{r}_i, \\ r^{int} &= \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\hat{r}_i - r^{ext})^2}, \\ r^{total} &= r^{ext} + \beta r^{int}, \end{aligned} \quad (3)$$

where r^{ext} is the mean value, r^{int} is the standard deviation, and r^{total} is the value that trade-off the exploration and exploitation using hyperparameter β. To perform online exploration in OGOE, we can infer the policy $\pi_{\theta'}(\cdot | prompt_{RL}(\mathbf{X}, X_{news}, r^{total}))$ for personalized news generation with exploration.

Algorithm 1 OGOE for Personalized News Headline Generation

Require: User behaviors $\mathbf{X}$, News X_{news}, SFT dataset D_{SFT}, RL dataset D_{RL}, Reward models $r_{\Phi_i}{}_{i=1}^{M}$, Goal target r_{target}.
Ensure: Personalized news headline $Y_{headline}$.

```
// Model Training
Train P_θ using D_SFT with Eq. (1).
Initialize π_θ' with θ_SFT.
Optimize π_θ' using D_RL.
// Exploration Phase
for i = 1 to M do
    Estimate reward r̂_i = r_Φi(X, X_news, Y_headline).
end for
Evaluate r^ext, r^int, and r^total.
if Exploration is needed then
    Adjust r_target = r^total.
end if
// Headline Generation
Produce Y_headline ~ π_θ'(·|prompt_RL(X, X_news, r = r_target)).
    return Y_headline
```

3.6 OGOE Framework

Figure 1 illustrates the architecture and workflow of personalized news headline generation. It comprises three parts:

Example Format (Figure 1(a)): Shows the process of generating a personalized news headline. The system uses previously clicked news headlines by the user, the current news content, a given prompt, and an expected personalized score as inputs to generate a personalized news headline suitable for the user.

Fine-tuning through Supervised Learning (Figure 1(b)): This part includes two steps. In the first step, using the PENS training set combined with previously clicked news by the user and the

current news along with a given prompt, GPT-3.5 is used to generate a batch of preliminary personalized news headlines. These generated headlines serve as the foundation for Supervised Fine-Tuning (SFT), teaching the model how to perform the task. Although these headlines might not score highly according to our target optimization strategy, the goal is to enable the model to grasp the basic process of the task, laying the groundwork for further fine-tuning. In the second step, we employ supervised learning to train the reward model, which is based on manually written personalized news headlines contained in the PENS test set. These data represent highly customized personalized news headlines, aiding us in developing a reward model capable of accurately assessing personalized news headline scores. During the training process, we randomly sample from the PENS dataset to prevent contamination of the training data with the test set.

Fine-tuning through RvS (Figure 1(c)): This section describes the fine-tuning process that combines Reinforcement Learning with Supervised Learning (RvS). In this process, the LLM takes personalized news headlines as input. Through these three independent evaluation processes, the calculation of the reward value becomes more reliable. Ultimately, the model uses these reward values to effectively generate personalized news headlines, reducing the need for frequent online fine-tuning and optimizing budget and resource utilization. The calculation of the reward model scores is detailed in section 4.1.3.

4 EXPERIMENT

To comprehensively assess our proposed method, we begin by evaluating the OGOE using the open-source PENS benchmark [5] and comparing it with other state-of-the-art methods. We then conduct an ablation study to demonstrate the generalization capability of the OGOE framework when given with various rewards, a feature not observed in prior work [21]. Lastly, we implement OGOE for the task of generating financial news headlines and it outperforms the existing online baseline methods.

4.1 Dataset, Base Model, and Reward Model

4.1.1 The PENS Dataset. The PENS dataset[1] [5] is a popular open-source benchmarks for research in personalized news headline generation with train and test sets. The training set contains about 113K English news articles with 15 categorical topics and 500K impression logs from over 445K anonymous users on Microsoft News. We perform evaluation using the test set which comprises over 100K personalized news headlines.

The PENS dataset provides a rich news browsing history for each user, depicting their long-term interest preferences. Specifically, each sample in the dataset contains the news articles clicked by the user prior to the current page exposure, including:

- The list of news IDs corresponds to the articles the user has clicked on in history.
- The dwell time the user spent on browsing those news.
- The timestamp when the user was exposed to those news.

We can leverage these data to model user interests. By employing a random sampling approach, we select 50 headlines from the news articles recently clicked by the user. These headlines are concatenated as contextual information for LLMs to represent user interest preferences, ensuring diversity in personalized features. For example, by constructing the recent news headlines clicked by a user:

Data Sample with $prompt_{RL}$

input:

There is an existing user,and the Headlines of the news he clicked are as follows: Headline 1: The national football team unveils its squad, featuring zhang yuning. Headline 2: A dip in china's GDP growth,setting at 2.2%. Headline 3: Tesla faces a staggering stock price crash, plummenting by over 30% globally. ... Headline 50: Recent research underscores the benefits of Omega-3 in thwarting heart diseases. The following news is available, [the source news article]. Please generate a personalized news headline for user on this news, and the Headline score is 1:

output:

Yao Ming predicts performance Chinese players in new NBA season.

This headline integrates the user's interest in sports news while incorporating domestic elements to enhance personalization. In our experiment, our model gives a reward score of 0.85 for this title, indicating a high match with the user's preferences.

4.1.2 BLOOMZ Model. In this study, we utilize BLOOMZ as personalized news headline generation base LLM. The BLOOMZ model was derived by researchers in the BigScience project through multitask fine-tuning built upon the BLOOM model [29]. The BLOOMZ model possesses the following characteristics:

(1) Built upon the BLOOMZ model, BLOOMZ is a large-scale multilingual pretrained language model obtained through pretraining on 46 billion language training tokens [40].
(2) Underwent multi-task prompted fine-tuning. The researchers constructed a multilingual multi-task training dataset, xP3, comprising 46 languages, and further fine-tuned BLOOMZ on this dataset to derive the BLOOMZ model.
(3) Exhibits strong performance on text-generation tasks. Results demonstrate that the BLOOMZ model surpasses the original BLOOMZ model on multiple text generation tasks, including translation and summarization.
(4) Scalable model sizes BLOOMZ models have versions ranging from 560M to 176B parameters to adapt different hardware.

Through multi-task-prompted fine-tuning on top of large-scale multilingual pretraining, the BLOOMZ model has obtained superior zero-shot generalization capabilities, especially in text generation. Therefore, BLOOMZ-1.1b [2] is chosen as the base model for the personalized news headline generation task. We perform the experiment on one NVIDIA GeForce RTX 3090 GPU which is a consumer-level GPU.

[1] https://msnews.github.io/pens.html

[2] https://huggingface.co/bigscience/bloomz-1b1

4.1.3 Reward Model. To ensure the effective headline generation, we design the reward function by considering the personalization, reality, and sensitivity. For personalization, our method utilize personalized reward function $r_{\text{personalized}}$, which is designed to align more effectively with user preferences. To ensure the veracity of content, we employ a reality function r_{reality} to evaluate and calibrate titles, thereby preserving the integrity of real news. For sensitivity, a sensitive word penalty $r_{\text{sensitive}}$ is integrated to mitigate the generation of titles with inappropriate language. Figure 2 displays sample headlines generated by our model alongside user rewritten one.

Figure 2: The graph delineates the shared lexicon between the personalized news headlines and the user-modified counterparts, alongside their respective average reward valuations.

These three reward values can be measures as follows:

- The personalization reward $r_{\text{personalized}} = \text{CosSim}(\boldsymbol{u}, \boldsymbol{h})$ assesses the degree of personalization in the generated headline by measuring the cosine similarity between the Sentence-Transformer embedding of user's previously clicked news titles $\boldsymbol{u}$ and the generated title $\boldsymbol{h}$.
- r_{reality}: Evaluates the fidelity of the generated headline to the factual content of the source article, utilizing ROUGE-1, ROUGE-2, and ROUGE-L metrics. The reality reward calculation can be expressed as:

$$r_{\text{reality}} = \text{ROUGE-1} + \text{ROUGE-2} + \text{ROUGE-L} \tag{4}$$

- $r_{\text{sensitive}}$: A binary value that represents the presence of inappropriate words from a lexicon of 1,000 offensive and sensitive terms, aiming to prevent inappropriate language in the headlines.

The overall reward function (r_{overall}) is formalized as the sum of each components, facilitating a balanced optimization for generating headlines that are personalized, factually accurate, and linguistically sensitive:

$$r_{\text{overall}} = r_{\text{personalized}} + r_{\text{reality}} - r_{\text{sensitive}}, \tag{5}$$

which is fitted by the reward models $\hat{r}_i$. In this experiment, we use 3 LoRA models [20] based on BLOOMZ-1.1b model to approximate the reward values due to it's parameter efficiency.

To train the reward models, we utilized 40% of the data from the PENS test set. These reward models were trained with the aforementioned reward functions r_{overall} serving as the target, optimizing the

Figure 3: Training loss curves for the reward model, indicative of the convergence of the three constituent reward functions.

LoRA parameters to fit the empirical reward values observed. Figure 3 illustrates the loss curves of these reward models throughout the training process.

4.2 Experiment Results

Here we perform the experiment on PENS dataset and evaluate the performance using reward model and typical metrics introduced in [5]. The hyperparameter we used is shown in Table 1.

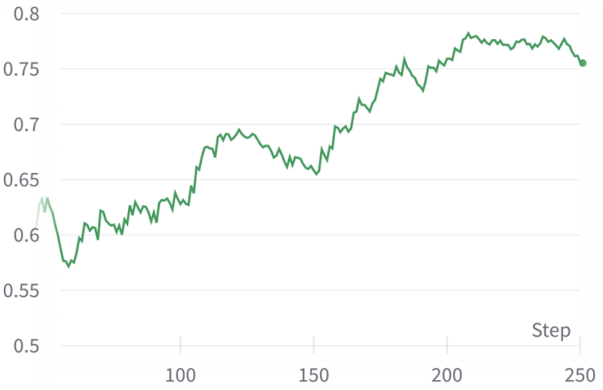

Figure 4: r^{ext} value of fine-tuning process which is evaluated on the test set with each training step.

4.2.1 Reward Evaluation of Personalized Headline Generation. We fine-tune the model using reward values as feedback under OGOE framework introduced in Section 3 to improve personalized headline generation on PENS dataset. During the test on test dataset, we set $r^{n+1} = 1$ to perform optimal personalized headline generation.

After fine-tuning by OGOE, the model is expected to improve at generating personalized headlines with higher reward evaluated by r^{ext}, which is introduced in Section 3.5. The value of reward feedback here demonstrate gradual learning and refinement, enhancing accuracy, personalization, and user satisfaction [24, 31, 43]. The

Table 1: Hyperparameters for OGOE Model

Hyper.	Sym.	Description	Default/Range	Impact/Notes
Click History Len.	N	Num. of user's clicked news titles	50	Personalization context
Personalization Score	X	Degree of personalization score	0.5	Adjusts personalization level
Model Params	θ	News generation model params	-	Determines performance
Policy Params	θ'	Policy model params in RvS	-	Dictates behavior under rewards
Expl.-Exploit. Trade-off	β	Balances exploration/exploitation	0.1	Balances strategy exploration
Target Reward	r_{target}	Target for headline generation	0.8	Encourages attractive headlines
Reward Models Num.	M	Independent reward models count	3	Improves reward estimation
Adam LR	α	Adam optimizer learning rate	1×10^{-5}	Affects training stability
CE Loss Weight	λ	Cross-entropy loss weight	1.0	Balances loss importance
News Content Len.	L	Length of news content	200	Affects detail level
Generated Title Len.	T	Length of generated news title	≤ 20	Ensures brevity

experiment results are shown in the Figure 4. We can observe that the reward evaluation is gradually increasing and reach around 0.77 score. Hence, we can conclude that the OGOE framework can indeed improve the personalized headline generation.

4.2.2 ROUGE Evaluation of Personalized Headline Generation. We evaluate generative quality using ROUGE-1, ROUGE-2, and ROUGE-L for informativeness and fluency [27] on test set. In this experiment, we choose Pointer-Gen [42], LSTUR [3], and PG+PL+ROUGE [52] as our baseline methods because these methods have been reported achieve state-of-the-art performance in headline generation evaluated by PENS.

Table 2: The overall performance of compared methods. "R-1, -2, -L" indicate F scores of ROUGE-1, -2, and -L,and "NA" denotes "Not Available". "IM" means injection methods, c.f. ①,②, and ③ refer to three injection methods derived from [5].

Methods	IM	Metrics		
		Rouge-1(%)	Rouge-2(%)	Rouge-L(%)
Pointer-Gen	NA	19.86	7.76	18.83
PG+PL+ROUGE	NA	20.56	8.42	20.03
LSTUR	①	23.71	8.73	**21.13**
LSTUR	②	24.1	8.82	20.73
LSTUR	③	23.11	8.42	20.38
OUR Model	NA	**25.03**	**8.97**	20.41

As shown in Table 2, the OGOE framework outperforms the state-of-the-art models evaluated by ROUGE-1 and ROUGE-2 metrics, demonstrating improved incorporation of key and personalized information into headlines [41]. Experimental results indicate that our framework can captures the principal content and ideas of articles more effectively, which is crucial for providing informative, comprehensive summaries encapsulating news essence. By accounting for individualized elements, our model can better match user interests and requirements when tailoring headlines. The experiment results also illustrate the personalization information in user view history is vital for improving headline quality and relevance.

4.2.3 Comparison with Various LLMs. In order to evaluate the effectiveness of our proposed method, we conducted a series of preliminary experiments comparing it against existing models, including Qwen[7], Llama2[50], and Yi[1]. These experiments were aimed at assessing the performance in the task of personalized headline generation. The experimental setup was consistent with the hyperparameters described in Table 1.

Table 3: Experimental results on personalized headline generation.

Model	Method	Rouge-1(%)	Rouge-2(%)	Rouge-3(%)
Qwen1.5-0.5B	Baseline	12.77	4.13	10.63
Qwen1.5-1.8B	Baseline	11.69	2.51	10.54
Qwen1.5-4B	Baseline	17.11	5.65	14.70
Yi-6B	Baseline	14.44	3.93	11.83
Bloomz-1.1B	Baseline	20.30	6.56	16.25
Bloomz-1.1B	Ours	25.03	8.97	20.41
Llama2-7B	Baseline	14.61	4.37	12.90
Llama2-7B	Ours	**26.31**	**9.26**	**20.98**

The results of these experiments are summarized in Table 3 below. As evident from the data, our method significantly enhances the performance across all ROUGE metrics compared to the baseline configurations of the models tested.

4.2.4 Generalization of Personalized Headline Generation through RvS. Additionally, we conduct experiment to investigate whether the LLMs trained under OGOE framework can generate personalized headlines aligned with specific reward values given in the $prompt_{RL}$ [12]. Specifically, we provide varying rewards $\hat{r^{n+1}} \sim \mathcal{U}[0, 1]$ with prompt $prompt_{RL}$ to produce corresponding personalized headlines on sampled data from test set. Then, we evaluate the generated outputs through reward model r^{ext}. Here, we plot the true input reward value and the output reward value in Figure 5. This evaluation is performed under distinct reward criteria [22].

We perform linear regression to plot the trend of predicted reward and ground truth. It confirms that the viability of our method for defining reward values and demonstrate the model's capability of adaptation in crafting reward-centric headlines. It also demonstrate that, by controlling the input reward signals, the model can actually generate different headlines with various personalized levels, content, and stylistic approaches.

Our research findings also provide strong support for the effectiveness of our proposed method in future investigations in LLMs generalization. The evaluations clearly illustrate that the LLMs have the ability to generate personalized titles based on specific interpolation rewards, while also highlighting possibilities for improving the accuracy and diversity of the results [43]. This observation is distinct from what has been commonly observed in previous studies that the goal-conditioned agents may experience difficulties in reward generalization.[34, 38].

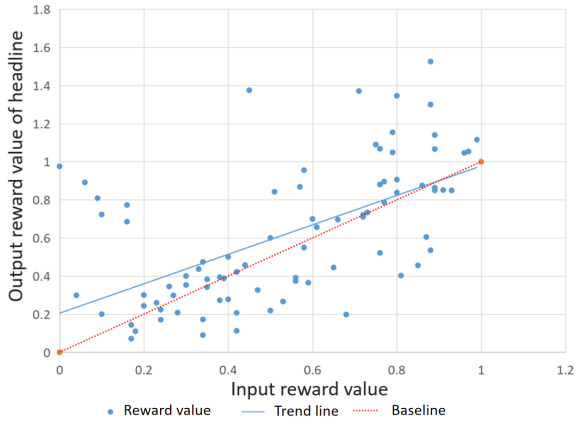

Figure 5: The points in the graph denote the reward values of 100 random samples from the PENS test set when generating personalized news headlines with designated reward values. The blue line signifies the trend line derived by applying least squares regression to the 100 sample points. The red line portrays a baseline at 45 degrees.

5 FINANCIAL REPORT HEADLINE GENERATION WITH ONLINE EXPLORATION

We have integrated our model, developed using the OGOE framework, into our financial report headline generation system. This system aims to create personalized headlines by considering the viewing history of financial analysts' reports and organizing important information based on their preferences. Initially, we trained the model using approximately 3000 data points. During the exploration phase, we updated the model three times using feedback from 200 data points in each phase, with a duration of one week per phase. To assess the performance of the model, we constructed a test set comprising 300 financial reports with labeled headlines and viewing history. Similar to the PENS experiment mentioned in Section 4, we evaluated the model's performance using ROUGE-1, ROUGE-2, and ROUGE-L metrics. As a baseline model for headline generation, we selected the SFT BLOOMZ-1b1 model, which was initially deployed online. The results, presented in Table 4, demonstrate a significant improvement in headline generation compared to the online base model, as observed in the experiment.

6 ABLATION

6.1 Reward Evaluation of OGOE, SFT, and BLOOMZ for Headline Generation

In the ablation study, we evaluate the performance of personalized headline generation of BLOOMZ itself and SFT model on PENS dataset, to clearly demonstrate the contribution of OGOE and BLOOMZ. For the experiment setup, we sampled 10,000 user data entries from the PENS test set. We used 20% (2,000 entries) as the test set, while the remaining 80% (8,000 entries) were utilized as the training set. For BLOOMZ, we directly evaluated the headline generation performance on the test set without any training or fine-tuning, to demonstrate inherent capabilities of our base model. In contrast, the SFT model θ_{SFT} was first performed fine-tuning on the training set and subsequently tested on the test set to gauge its performance of directly aligning. Here, we trained SFT model with learning rate of 0.00001, under cross-entropy loss, and utilize the Adam optimizer. Similar to the measurement used in Section 4.2.1, we evaluate the output by reward through r^{ext}. The results are shown in Figure 6.

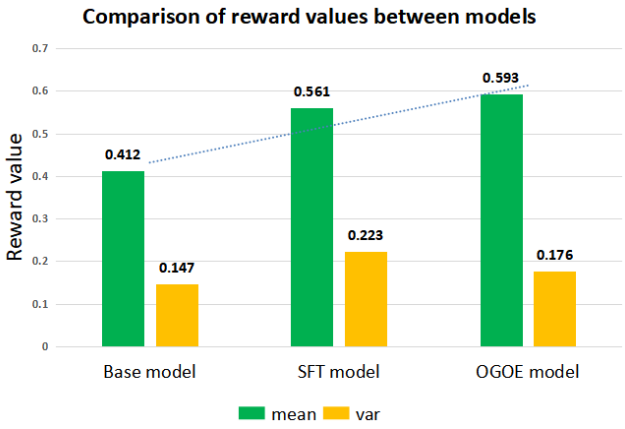

Figure 6: The mean and variance of personalized news headline generation evaluated by r^{ext}. Here we plot the base model θ, the SFT model θ_{SFT}, and the final model θ' after training under OGOE framework.

The results show the superiority of our model θ' trained under OGOE framework over both the base BLOOMZ model θ and SFT model θ_{SFT}, emphasizing the advantages of OGOE in incorporating personalization and generate headlines with RvS techniques. Our approach's superiority over the baseline also highlight the significance of utilize personalized factors, such as user preferences and context, in customizing headlines to individual needs [26]. Moreover, the incremental performance compared to SFT model θ_{SFT} highlights the significance of fine-tuning on personalized data with goals [19]. This iterative process allows our model to adapt and refine, substantially boosting performance for personalized headline generation [45].

Table 4: Comparison of models using few-shot metrics (percentages).

Metrics (Few-Shot)	ROUGE-1 (%)			ROUGE-2 (%)			ROUGE-L (%)		
Models	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall
BASE	20.54	18.31	34.99	12.27	9.98	25.37	18.38	15.41	34.77
OURS	**54.56**	**50.82**	**62.39**	**39.28**	**35.65**	**46.70**	**51.60**	**47.52**	**62.43**

6.2 The Impact of Leveraging Reward Learning for Personalized, Reality and Sensitive-Aware Headline Generation

In the experiment of Section 4, we utilize the sum value of three individual reward values. In this ablation, we observe the capability of reward models in modeling personalization, reality, and sensitivity, by observing the value separately. It is worth to know that, these three perspectives are the most common optimization direction of headline generation. We evaluate the reward on test set and the experiment result is shown in the Figure 7. Compared to the baseline, our method demonstrates superior performance in generating personalized news headlines across all the settings. The OGOE can reduce the generation of sensitive content, improve the content reality, and enhance the personality of the headline, respectively.

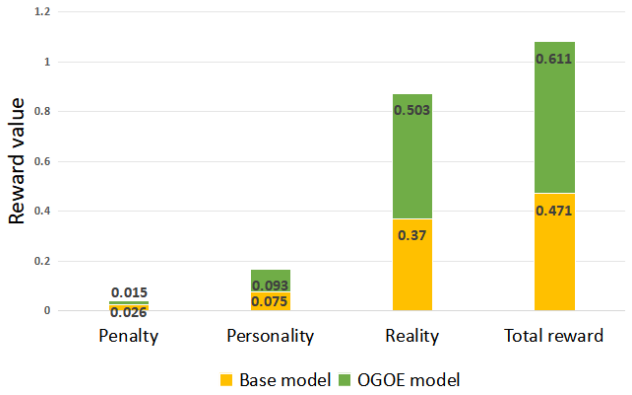

Figure 7: Reward values corresponding to the base model and the model trained by OGOE evaluated by the sensitive word penalty reward function, personalized reward function, and reality reward function.

7 CONCLUSION

In this paper, we propose a novel framework OGOE for personalized news headline generation, utilizing offline goal-conditioned reinforcement learning to perform online exploration with LLMs. This approach integrates user information and behavior into a natural language prompt, performs goal-conditioned reinforcement learning via SFT, and performs online exploration by only modifying the goals without frequently retraining the LLMs. The results from our extensive evaluation, conducted on publicly available benchmarks, custom simulations, and real-world financial report headline generation, showcase the superior performance and robust generalization capability of LLMs under OGOE. This study marks a pioneering effort in assessing the offline goal-conditioned RL capabilities of LLMs and underscores a new effective approach for personalized headline generation, further expanding the practical utility and theoretical understanding of LLMs in RL scenarios.

8 ACKNOWLEDGE

This work is supported by the National Natural Science Foundation of China (62102241) and Shanghai Municipal Natural Science Foundation (23ZR1425400).

REFERENCES

[1] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652 [cs.CL]

[2] Enrique Alfonseca, Daniele Pighin, and Guillermo Garrido. 2013. Heady: News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1243–1253.

[3] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 336–345. https://doi.org/10.18653/v1/P19-1033

[4] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. 2020. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990* (2020).

[5] Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline 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)*. 82–92.

[6] Ayana, Shi-Qi Shen, Yan-Kai Lin, Cun-Chao Tu, Yu Zhao, Zhi-Yuan Liu, and Mao-Song Sun. 2017. Recent advances on neural headline generation. *Journal of computer science and technology* 32 (2017), 768–784.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv:2309.16609 [cs.CL]

[8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter,

Chris 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 *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[10] Nicolas Carrara, Edouard Leurent, Romain Laroche, Tanguy Urvoy, Odalric-Ambrym Maillard, and Olivier Pietquin. 2019. Budgeted reinforcement learning in continuous state space. *Advances in Neural Information Processing Systems* 32 (2019).

[11] Eric Chen, Zhang-Wei Hong, Joni Pajarinen, and Pulkit Agrawal. 2022. Redeeming intrinsic rewards via constrained optimization. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 4996–5008. https://proceedings.neurips.cc/paper_files/paper/2022/file/204fee94c982a19230c39045aa54f977-Paper-Conference.pdf

[12] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[13] Corinna Cortes, N Lawarence, D Lee, M Sugiyama, and R Garnett. 2015. Advances in neural information processing systems 28. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*.

[14] Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*. 1–8.

[15] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine* 39, 3 (2022), 42–62.

[16] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 11737–11762. https://doi.org/10.18653/v1/2023.acl-long.656

[17] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[18] Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*. Springer, 87–93.

[19] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[21] Yongkeun Hwang, Hyeongu Yun, and Kyomin Jung. 2021. Contrastive Learning for Context-aware Neural Machine Translation Using Coreference Information. In *Proceedings of the Sixth Conference on Machine Translation*, Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (Eds.). Association for Computational Linguistics, Online, 1135–1144. https://aclanthology.org/2021.wmt-1.121

[22] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems* 31 (2018).

[23] Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4037–4058.

[24] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[27] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[28] Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: Human reading inspired extractive summarization. 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)*. 3033–3043.

[29] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786* (2022).

[30] Kazuma Murao, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A case study on neural headline generation for editing support. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. 73–82.

[31] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. Citeseer, 278–287.

[32] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[34] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* (2017).

[35] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000* (2018).

[36] Xihe Qiu, Teqi Hao, Shaojie Shi, Xiaoyu Tan, and Yu-Jie Xiong. 2024. Chain-of-LoRA: Enhancing the Instruction Fine-Tuning Performance of Low-Rank Adaptation on Diverse Instruction Set. *IEEE Signal Processing Letters* 31 (2024), 875–879. https://doi.org/10.1109/LSP.2024.3377590

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[38] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. Publisher Copyright: © ICLR 2016: San Juan, Puerto Rico. All Rights Reserved.; 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.

[39] Gustavo Sandoval, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt. 2023. Lost at c: A user study on the security implications of large language model code assistants. *arXiv preprint arXiv:2208.09727* (2023).

[40] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[41] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[42] 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)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1073–1083. https://doi.org/10.18653/v1/P17-1099

[43] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[44] Rui Sun, Yue Zhang, Meishan Zhang, and Donghong Ji. 2015. Event-driven headline generation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 462–472.

[45] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355* (2019).

[46] Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1054–1059.

[47] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503* (2021).

[48] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach.. In *IJCAI*, Vol. 17. 4109–4115.

[49] Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-Criticism: Aligning Large Language Models with their Understanding of Helpfulness, Honesty, and Harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Mingxuan Wang and Imed Zitouni (Eds.). Association for Computational Linguistics, Singapore, 650–662. https://doi.org/10.18653/v1/2023.emnlp-industry.62

[50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[52] Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3065–3075. https://doi.org/10.18653/v1/D19-1303

[53] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* 31 (2018).