

# Online Personalizing White-box LLMs Generation with Neural Bandits

Zekai Chen  
J.P. Morgan Chase  
New York, NY

zekai.chen@jpmchase.com

Weeden Daniel Po-yu Chen Francois Buet-Golfouse  
J.P. Morgan Chase  
London, UK

francois.buet-golfouse@jpmorgan.com

## Abstract

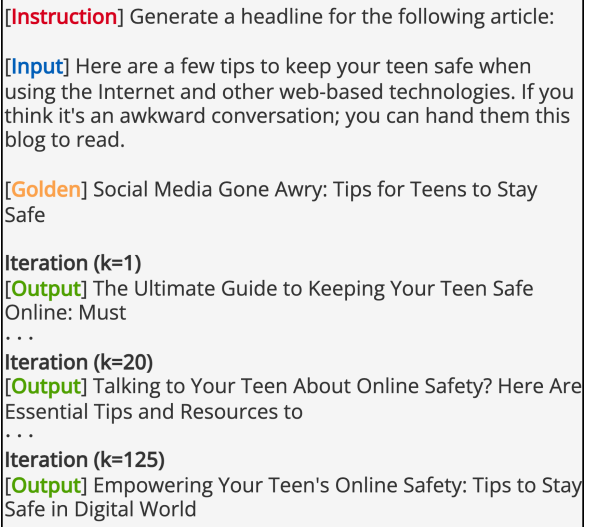
The advent of personalized content generation by LLMs presents a novel challenge: how to efficiently adapt text to meet individual preferences without the unsustainable demand of creating a unique model for each user. This study introduces an innovative online method that employs neural bandit algorithms to dynamically optimize soft instruction embeddings based on user feedback, enhancing the personalization of open-ended text generation by white-box LLMs. Through rigorous experimentation on various tasks, we demonstrate significant performance improvements over baseline strategies. NeuralTS, in particular, leads to substantial enhancements in personalized news headline generation, achieving up to a 62.9% improvement in terms of best ROUGE scores and up to 2.76% increase in LLM-agent evaluation against the baseline.

## 1 Introduction

In recent years, the advancements in large language models (LLMs) have been remarkable (Brown et al., 2020; Zhao et al., 2023), with these models demonstrating an unparalleled ability to understand and generate text across a wide spectrum of tasks (Wei et al., 2022; Kojima et al., 2022). This capability has revolutionized the way we interact with machine-generated content and opened up new avenues for personalized text generation (Kirk et al., 2023; Li et al., 2023a).

Personalization in text generation is of paramount importance to ensure user engagement and satisfaction (Huang et al., 2022) across a range of applications such as composing tweets, or generating news articles and financial reports, or in more personalized settings like business communications and creative writing (Li and Tuzhilin, 2019; Li et al., 2020).

However, the prospect of developing a unique LLM for each user presents challenges, including the prohibitive resource requirements (Hoff-



```
[Instruction] Generate a headline for the following article:

[Input] Here are a few tips to keep your teen safe when
using the Internet and other web-based technologies. If you
think it's an awkward conversation; you can hand them this
blog to read.

[Golden] Social Media Gone Awry: Tips for Teens to Stay
Safe

Iteration (k=1)
[Output] The Ultimate Guide to Keeping Your Teen Safe
Online: Must
...
Iteration (k=20)
[Output] Talking to Your Teen About Online Safety? Here Are
Essential Tips and Resources to
...
Iteration (k=125)
[Output] Empowering Your Teen's Online Safety: Tips to Stay
Safe in Digital World
```

Figure 1: Evolution of generated headlines for an article on teen internet safety, illustrating the progressive refinement of generation that emulates this journalist stylistic tendencies through online learning.

mann et al., 2022), data privacy concerns (Li et al., 2023b), and the scarcity of personalized data (Rafailov et al., 2023). These obstacles necessitate an alternative strategy that is both practical and flexible. A promising solution lies in adopting lightweight models capable of online learning, which can dynamically adjust their output based on continuous user feedback (Bai et al., 2022). Such an approach not only circumvents the need for a bespoke model for each user but also encourages alignment of the generated content to individual preferences over time. Importantly, this adaptive process is poised to unlock long-term rewards stemming from personalization, encompassing not just explicit preferences expressed by users but also responding to favorable actions (Xie et al., 2021).

Despite these benefits, the ultimate effectiveness of LLMs hinges on the quality of the given instructions (Zhou et al., 2022; Bang et al., 2023; White et al., 2023). Previous efforts have focused on gradient-based strategies (Shin et al., 2020; Li

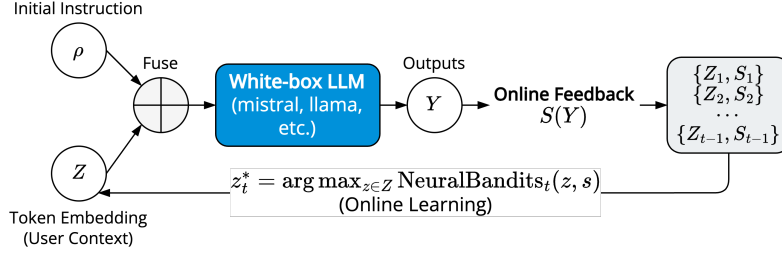


Figure 2: Illustration of our framework. Details are described in Section 2.

and Liang, 2021; Lester et al., 2021) for automated instruction optimization, the applicability is limited to less advanced public models, leaving out many advanced yet proprietary models. With the emergence of more advanced open models such as Mistral-7B (Jiang et al., 2023), Llama-70B (Touvron et al., 2023a,b), and Mixtral-8x7B (Jiang et al., 2024), which offer transparency and have reported performance that even surpasses that of ChatGPT-3.5<sup>1</sup>, there is a renewed focus on leveraging these models for direct optimization.

In this study, we introduce a novel online method for enhancing the personalization of open-ended text generation with white-box LLMs. Considering that *capturing the nuances of persona in natural language instructions is challenging*, we aim to directly optimize the soft token embeddings (Chen et al., 2023; Lin et al., 2023), representing the contextual factors through user feedback by utilizing neural bandit algorithms (Zhou et al., 2019; Zhang et al., 2021). This method not only promises to refine the personalization process of text generation but also contributes to the broader application of adaptive algorithms in creating content that closely reflects individual user preferences.

## 2 Personalization with Neural Bandits

Neural Bandits (Zhou et al., 2019; Zhang et al., 2021) integrate the adaptive exploration of bandit algorithms (Auer, 2003; Agrawal and Goyal, 2012; Li et al., 2010) with neural networks’ superior ability to predict rewards under uncertainty. By leveraging past interactions to balance the trade-off between exploring new actions and exploiting known ones, these algorithms can accurately predict and enhance personalized outcomes. Therefore, we adopt NeuralUCB (Zhou et al., 2020) and NeuralTS (Zhang et al., 2021) in our framework to comprehensively evaluate how Neural Bandits benefit LLMs’ generation in an online fashion.

This section explores the application of Neural Bandits to white-box large language models (LLMs) as a strategy for directly refining soft prompts (*aka.* contextual embeddings) to overcome the *inability of natural language instructions to fully express nuances of persona*. The process involves a white-box LLM,  $f$ , which takes a soft prompt  $z$ , an instruction,  $\rho$ , and a test input,  $x$ , to produce an output sentence,  $\hat{y} = f(z, \rho, x)$ . A soft prompt  $z$  is a continuous vector representing preference token embeddings, which is used alongside the default instruction,  $\rho$ , as input to  $f$ . Given a specific user  $u$  at time  $t$ , we have a specific input  $x_i$ , the goal is to find an optimal soft prompt token,  $z^*$ , that maximizes the following objective function based on user’s feedback  $s(\hat{y}_i)$  (see Figure 2 for illustration):

$$z_t^* = \arg \max_{z \in Z} \text{NeuralBandits}_t(z, s)$$

With NeuralUCB, it involves computing the acquisition value for each candidate soft prompt and selecting the one that maximizes this value, based on the model’s current parameters and the uncertainty associated with each prompt:

$$z_{t+1} = \arg \max_{z \in Z} (\mu(g(z); \theta_t) + \nu_t \sigma_t(g(z); \theta_t)),$$

Here,  $\mu(g(z); \theta_t)$  represents the network’s predicted value for the soft prompt  $z$  at iteration  $t$ , while  $\sigma_t(g(z); \theta_t)$  quantifies the uncertainty of the prediction. The parameter  $\nu_t$  controls the trade-off between exploration and exploitation, influencing the algorithm’s preference for exploring less certain prompts versus exploiting prompts with higher predicted values.

When using NeuralTS for updating  $z$ , the process involves sampling from a predictive distribution to select the next soft prompt. Unlike NeuralUCB, which directly uses a deterministic acquisition function, NeuralTS generates a sample for each candidate  $z$  from a distribution modeled by the neural network. The update can be represented as follows:

<sup>1</sup><https://chat.openai.com/>

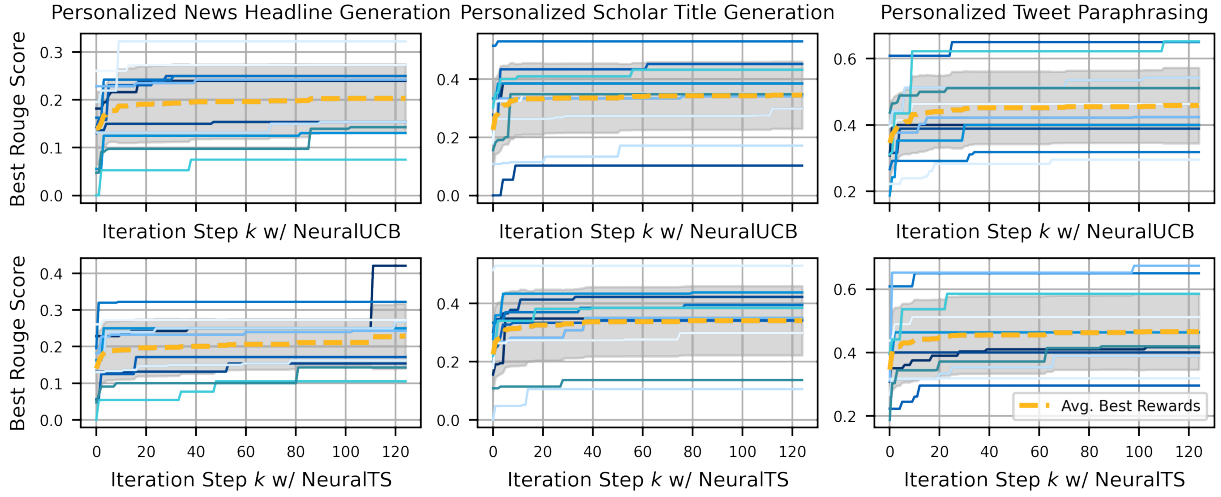


Figure 3: 10 user profiles (different blues) are randomly selected for demonstration. Trend of increasing averaged best rewards (yellow dashes) across learning iterations for three personalized text generation tasks, showcasing the progressive improvement in performance achieved by both NeuralUCB (Zhou et al., 2019) and NeuralTS (Zhang et al., 2021) algorithms.

$$z_{t+1} = \arg \max_{z \in Z} \tilde{r}_{z,t}$$

$$\tilde{r}_{z,t} \sim \mathcal{N}\left(\tilde{\mu}(g(z); \theta_t), \nu_t \tilde{\sigma}_t(g(z); \theta_t)\right)$$

Where,  $\tilde{\mu}(g(z); \theta_t)$  and  $\tilde{\sigma}_t(g(z); \theta_t)$  represent the estimated mean and standard deviation respectively for the soft prompt  $z$ , and  $\tilde{r}_{z,t}$  is a sampled reward from the predictive distribution for  $z$  at iteration  $t$ . The parameter  $\nu_t$  again balances exploration and exploitation, but in the context of NeuralTS, the exploration is informed by the stochasticity introduced through sampling, encouraging diversity in the selection of  $z$  based on both prediction and uncertainty.

NeuralUCB uses an upper confidence bound to balance these aspects deterministically, offering robust performance in environments where a clear quantification of uncertainty benefits decision-making. NeuralTS, on the other hand, employs a Thompson sampling (Agrawal and Goyal, 2012) approach, introducing stochasticity in the selection process, which can lead to more diverse exploration. Therefore, we tested both approaches as different kernels given the actual performance hinges on the problem’s nature, the desired balance between exploration and exploitation, and the computational resources available.

### 3 Experiments on LaMP

#### 3.1 Personalized Generations

The LaMP (Salemi et al., 2023) dataset, or Language Model Personalization, is a benchmark de-

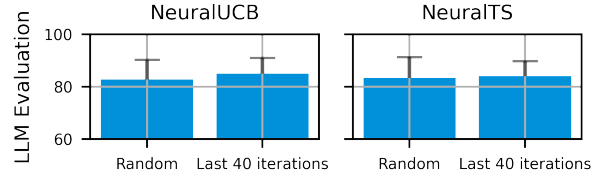


Figure 4: LLM evaluation of personalized generation between NeuralUCB (Zhou et al., 2019) and NeuralTS (Zhang et al., 2021) in personalized news headline generation.

signed for training and evaluating large language models (LLMs) to produce personalized outputs. It aims to assess the efficacy of LLMs in generating responses tailored to individual user profiles. For example, in personalized news headline generation task, the writing of journalist exhibits unique stylistic characteristics shaped by both individual and societal influences (Zhu and Jurgens, 2021). This scenario serves as an excellent opportunity for exploring personalized generation.

In this study, we evaluate our framework on three open-ended generation tasks which are Personalized News Headline Generation, Personalized Scholarly Title Generation, and Personalized Tweet Paraphrasing.

#### 3.2 Online Simulation using LaMP

We simulate a realistic online setting where individual users, such as those represented by distinct profiles in the LaMP dataset, continually receive outputs from LLMs in response to predetermined instructions. We then take ROUGE scores (Lin, 2004), which measure the correspondence between

Tasks	Metric	Random (zero-shot)	NeuralUCB @ $k = 165$	NeuralTS @ $k = 165$	$\Delta \uparrow$
Personalized News Headline Generation	Avg. ROUGE-1/L $\uparrow$	$0.140 \pm 0.076$	$0.203 \pm 0.072$	<b><math>0.228 \pm 0.088</math></b>	62.9%
Personalized Scholarly Title Generation	Avg. ROUGE-1/L $\uparrow$	$0.225 \pm 0.127$	<b><math>0.345 \pm 0.116</math></b>	$0.341 \pm 0.119$	53.3%
Personalized Tweet Paraphrasing	Avg. ROUGE-1/L $\uparrow$	$0.346 \pm 0.110$	$0.459 \pm 0.114$	<b><math>0.466 \pm 0.120</math></b>	34.7%

Table 1: Performance comparison of NeuralUCB (Zhou et al., 2019) and NeuralTS (Zhang et al., 2021) algorithms against a random baseline in three personalized text generation tasks, measured by average ROUGE-1/L scores with improvement percentages ( $\Delta \uparrow$ ) after 165 iterations.

the model’s output and the user’s ideal or "golden" response, as their *online feedback*. Additionally, we monitor the assessments from a black-box LLM, which evaluates the appropriateness of the generated content for a desired persona based on their profiles (*e.g.*, using LLMs to automatically rate the generation based on the consistency with users’ stylistics). An ideal result would exhibit both improved ROUGE metrics and LLM ratings.

As introduced in Section 1, we employ Mistral-7B-Instruct-v0.2<sup>2</sup> as our default white-box LLM. By simulating this online interaction scenario, our aim is to closely mimic the dynamic and personalized experience users encounter in real-world applications of language models. Through this experiment, we seek to evaluate the effectiveness of Neural Bandits in responding to the nuanced persona of different users/profiles, thereby informing future improvements and adaptations in personalized generation.

### 3.3 Results Analysis

For the task of Personalized News Headline Generation, from Table 1 we observed a significant performance leap with NeuralTS, achieving an Average ROUGE-1/L score of  $0.228 \pm 0.088$ , marking a 62.9% improvement over the random baseline ( $0.140 \pm 0.076$ ). NeuralUCB also outperformed the baseline, albeit with a slightly lower score of  $0.203 \pm 0.072$ . In Personalized Scholarly Title Generation, NeuralUCB exhibited the highest increase, with a score of  $0.345 \pm 0.116$ , closely followed by NeuralTS ( $0.341 \pm 0.119$ ), both surpassing the baseline ( $0.225 \pm 0.127$ ) by approximately 53.3%. Lastly, the Personalized Tweet Paraphrasing task showed NeuralTS slightly outperforming NeuralUCB with a score of  $0.466 \pm 0.120$  against  $0.459 \pm 0.114$ , over a baseline of  $0.346 \pm 0.110$ ,

translating to a 34.7% improvement. Figure 2 also summarises these results. At the same time, as shown in Figure 4, we observe increase over the LLM evaluation scores, with NeuralUCB achieving an average improvement of 2.8% vs.  $\sim 1\%$  for NeuralTS. These findings underscore the potential of Neural Bandit algorithms in enhancing content personalization across various text generation tasks.

However, it’s important to acknowledge certain limitations inherent in LLM evaluations. LLMs tend to exhibit a positivity bias, often returning higher scores for the generated content. This bias can potentially skew the evaluation results and may not fully capture nuances in personalization or consistency. Also, while ROUGE provides a useful automated proxy for assessing the quality of personalized content, it has well-known limitations in fully capturing subjective preferences or nuanced stylistic tendencies. It may not align well with human judgments, especially for highly creative and personalized text. Thus, while the gains demonstrated in both ROUGE and LLM evaluation provide encouraging evidence, human evaluations are imperative for comprehensively assessing improvements in adapting to individual user profiles.

## 4 Conclusion

In conclusion, this study presents a novel framework for personalized content generation using large language models (LLMs) through the application of neural bandit algorithms. By dynamically optimizing soft instruction embeddings through online feedback, our method demonstrates improvements in personalizing open-ended text generation tasks as scored by both ROUGE scores and an LLM evaluator. Further research is needed to confirm that these results are consistent with human-evaluated suitability.

<sup>2</sup><https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>



## 5 Limitations

We discuss some limitations this work may present in this section. The evaluation in this study is limited to only three generation tasks from the LaMP benchmark. Testing on a more diverse and extensive set of personalized tasks would strengthen the claims around the general applicability of the method. The simulation of an online environment with user profiles from the LaMP dataset may not fully capture the nuances of real-world user interactions and feedback. Evaluating with actual human subjects over longer time horizons could reveal additional challenges. Additionally, ROUGE score, is used to assess generation quality. Incorporating other metrics and human evaluations could provide a more holistic view of the improvements achieved. As of now, only two neural bandit algorithms, NeuralUCB and NeuralTS, are investigated. Comparing to a wider range of adaptive optimization algorithms could provide deeper insights into the most suitable techniques.

## 6 Ethics Consideration

In real-world, collecting personalized data like past personal data raises privacy concerns around data usage and consent. In our study, we use LaMP which is a public dataset with user consent. To some extent, tailoring content to align with a user's preferences risks creating isolated filter bubbles that entrench viewpoints. This is also an important reason that we investigate bandit algorithms to better balance exploration against exploitation and serve as mitigation strategies which exposes users to diverse perspectives. One potential risk is that the capability to generate highly persuasive personalized text could be misused for deceptive purposes. Safeguards against generating harmful, unethical or untruthful content need to be in place. Like any powerful technology, personalized generation abilities could be co-opted for nefarious ends by malicious actors. Responsible disclosure and governance practices are critical. In summary, the ethical dimensions span privacy, security, fairness, transparency, accountability and the broader social impacts of deploying personalized generative models. A principled, human-centric approach that places ethics at the foundation of research and development is imperative as these capabilities continue to advance.

## References

- Shipra Agrawal and Navin Goyal. 2012. [Thompson sampling for contextual bandits with linear payoffs](#). In *ICML*.
- Peter Auer. 2003. [Using confidence bounds for exploitation-exploration trade-offs](#). *JMLR*, 3:397–422.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *ArXiv*, abs/2204.05862.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity](#). *IJNLP*, abs/2302.04023.
- 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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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](#). *NeurIPS*, abs/2005.14165.
- Lichang Chen, Jiuhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou. 2023. [Instructzero: Efficient instruction optimization for black-box large language models](#). *ArXiv*, abs/2306.03082.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. [Training compute-optimal large language models](#). *ArXiv*, abs/2203.15556.
- Xiaolei Huang, Lucie Flek, Franck Dernoncourt, Charles F Welch, Silvio Amir, Ramit Sawhney, and Diyi Yang. 2022. [UserNLP'22: 2022 international workshop on user-centered natural language processing](#). *Companion Proceedings of the Web Conference 2022*.

- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. [Mixture of experts](#). *ArXiv*, abs/2401.04088.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. [Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback](#). *ArXiv*, abs/2303.05453.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). *NeurIPS*, abs/2205.11916.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *EMNLP*.
- Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023a. [Teach llms to personalize - an approach inspired by writing education](#). *ArXiv*, abs/2308.07968.
- Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, and Yangqiu Song. 2023b. [Multi-step jailbreaking privacy attacks on chatgpt](#). *EMNLP*, abs/2304.05197.
- Junyi Li, Siqing Li, Wayne Xin Zhao, Gaole He, Zhicheng Wei, Nicholas Jing Yuan, and Ji rong Wen. 2020. [Knowledge-enhanced personalized review generation with capsule graph neural network](#). *CIKM*.
- Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. [A contextual-bandit approach to personalized news article recommendation](#). In *The Web Conference*.
- P. Li and Alexander Tuzhilin. 2019. [Towards controllable and personalized review generation](#). In *EMNLP*.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). *ACL*, abs/2101.00190.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *ACL*.
- Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. 2023. [Use your instinct: Instruction optimization using neural bandits coupled with transformers](#). *ArXiv*, abs/2310.02905.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). In *ICLR*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *ArXiv*, abs/2305.18290.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. [LaMP: When large language models meet personalization](#).
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. [Eliciting knowledge from language models using automatically generated prompts](#). *EMNLP*, abs/2010.15980.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv*, abs/2307.09288.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *NeurIPS*, abs/2201.11903.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. [A](#)

prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*, abs/2302.11382.

Tengyang Xie, John Langford, Paul Mineiro, and Ida Momennejad. 2021. [Interaction-grounded learning](#). In *ICML*.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. 2021. Neural thompson sampling. In *ICLR*.

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, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. [A survey of large language models](#). *ArXiv*, abs/2303.18223.

Dongruo Zhou, Lihong Li, and Quanquan Gu. 2019. [Neural contextual bandits with ucb-based exploration](#). In *ICML*.

Dongruo Zhou, Lihong Li, and Quanquan Gu. 2020. Neural contextual bandits with ucb-based exploration. In *ICML*, pages 11492–11502. PMLR.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. [Large language models are human-level prompt engineers](#). *ICLR*, abs/2211.01910.

Jian Zhu and David Jurgens. 2021. [Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles](#). In *EMNLP*.

## A Appendix

### A.1 Hyperparameters for NeuralBandits

Following InstructZero (Chen et al., 2023) and Instinct (Lin et al., 2023), the reported results were based on hyperparameters optimized as:

- Intrinsic dimension - The soft prompt  $z$  often has very high dimensionality (e.g.  $d = 4096 \times N_z$  for Mistral-7B), making it challenging to directly optimize via Bayesian optimization. To address this, InstructZero (Chen et al., 2023) employs random projection as an effective dimensionality reduction technique. Specifically, a random projection matrix  $A \in \mathcal{R}^{d \times d'}$  is used, where  $d' \ll d$  is the reduced intrinsic dimension. Given a  $d'$ -dimensional vector  $z'$  and the projection matrix  $A$ , the original high-dimensional prompt  $z$  is computed as  $z = Az'$ . By optimizing over  $z'$  instead of  $z$ , the dimensionality of the optimization problem is reduced from  $d$  to the much lower  $d'$ .  $d'$  is known as the intrinsic dimension in this case. We set it as 100 by default usage.

- Number of soft tokens -  $N_z$  represents the length of contextual embeddings or soft tokens that

are utilized to concatenate with initial instructions fed into LLMs. Ideally, we would like to balance off this parameter so it's able to capture the persona nuances while also not causing extra compute redundancy. This number is set to 5 in our experiments.

- $\lambda$  - Controls the strength of the prior in the NeuralUCB acquisition function. Set to 0.1 after grid search.

- $\nu_t$  - balances exploration versus exploitation in Neural Bandits. Fixed at 0.1 based on best validation performance.

- Total iterations - total update iterations for neural bandits, set to 165 in practice.

- Hidden dimensions - We used an MLP on top of the LLM embeddings with 100 hidden units. This architecture achieved the best validation results.

- Local iterations - represents how many learning steps for neural network to retrain every update iteration. In order to prevent overfitting but also avoid underfitting, we set this number as 40 in practice.

- Learning rate - The MLP was trained with an AdamW optimizer (Loshchilov and Hutter, 2017) using a learning rate of  $3e^{-4}$ .

- Projection matrix - Random projections for dimensionality reduction were generated by sampling from a Uniform( $-1, 1$ ) distribution.

For simplicity, we randomly select 100 profiles from each task for experiments and all the reported results are average with standard deviation using the same seed 42. All experiments are conducted on 4 Nvidia A10g GPUs.

### A.2 Prompts for LLM Evaluation

For LLM-agent evaluation, we design prompts for black-box LLM that evaluates the white-box LLMs generation in a more comprehensive way. Please refer to Figure 4 for the details.

You are a journalist who produces news headline with your own writing stylistic. Here are samples you generated in the past: '[**demos**]'.  
  
Now, we have the following article: '[**article**]'.  
  
What would you rate this title '[**title**]' on a scale of 1 to 100 based on the consistency with your style and how faithfully it represents the original content? Be very rigorous and judge with a basis.

Figure 5: Using personalized news headline generation as an example. Prompts fed to the black-box LLMs for human-like evaluation of the generation by white-box LLM.