Push for New Users and New Content: Overview of Content Generation and Personalization of Large Language Model in Cold Start Scenario

Ziang Li*

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

Abstract. Push system plays an important role in personalized services of various platforms, but often due to the lack of new user behavior data and interactive data of new content, resulting in the cold start problem, the traditional system cannot accurately push such new users. In recent years, the rapid development of large language model (LLM) shows its advantages in text rich and data sparse environments by using the transmission based paradigm, semantic understanding and pre training knowledge. This paper systematically summarizes the core methods of LLM in new user portrait reasoning, push copy generation, new content understanding and push scheme construction. The technical paths based on cue engineering, context learning, retrieval enhanced generation (RAG) and meta learning are analyzed respectively, and the advantages and disadvantages of pure LLM and hybrid architecture are compared. Finally, this paper summarizes the four challenges of hallucination control, reasoning delay, privacy ethics and effect evaluation, and looks forward to the future direction of the combination of LLM and push system, so as to provide reference for the construction of the next generation of recommendation system with better effect.

1 Introduction

As a means to attract new users and increase user retention, push notifications have been valued by major platforms as standard capabilities. The effect of push is highly related to the degree of personalization. Users will be attracted to the information they are interested in, but one of the problems of personalized push has always been the cold start problem. In the push environment, this problem can be further refined into two different forms. One is that new users do not leave any behavior data on the platform, and the other is that no new content generates any interactive data. These two forms often occur together and overlap each other, making the cold start problem more difficult.

The traditional method is essentially a kind of "behavior matching". In the case of cold start, the matching logic cannot be established due to the lack of data. Moreover, the traditional method does not have the ability of semantic generation, and cannot provide personalized content for new users at the first contact. This is laying the groundwork for the paradigm breakthrough of LLM.

The rise of large language model provides a new solution to the cold start problem [1]. Compared with the traditional methods, the advantages of the large language model are reflected in its strong semantic understanding ability and pre training knowledge, so that it can infer the potential preferences of users from the information such as device type and installation location in the absence of user data. In addition, the powerful language generation ability of the large language model enables it to generate corresponding and customized push copies for different new users in real time, breaking away from the shackles of the traditional fixed template. This means that the transformation of push from "passive matching based on behavior" to "active understanding and generation based on semantics" just helps to solve the two problems of understanding who new users are and accurately providing new content that need to be solved urgently in the cold start problem.

In recent years, many directions of research have begun to explore LLM in solving the problem of push intercooling startup. First of all, there is an industrial level LLM push generation landing. Pushgen is deployed on a large scale in the fast hand. By using the controllable category prompt and reward model sorting, it realizes the personalized push document generation with the same effect as manual writing, which directly verifies the engineering feasibility of LLM in the push notification scenario [2]. Secondly, the recommendation ability of LLM under the condition of lack of data has also been proved. The system comparison experiment shows that the performance of LLM under the condition of zero samples/few samples is comparable to that of traditional collaborative filtering [3]. In addition, meta learning has greatly enhanced LLM's ability to quickly adapt to new users [4]. The above work has made great progress in all directions, but there are still blind spots. First of all, most studies have not been tested under the constraints of the high real-time nature of the actual push

*Ziang.Li24@student.xjtlu.edu.cn

and the fatigue of users. Second, new users and new content are often confused. At present, there is still a lack of a review that integrates the two into a unified framework and systematically reviews them in combination with the progress of LLM technology.

LLM has unique advantages in the environment of rich text and sparse data [1]. This article will focus on the content generation and personalization of LLM in the push cold start scenario, and systematically sort out the technical context, compare the advantages and disadvantages of the route, identify the key challenges and point out the future direction from the dual dimensions of new users and new items

The structure of subsequent articles is as follows: Section 2 reviews basic technologies, section 3 analyzes two types of applications, section 4 discusses challenges and prospects, and section 5 summarizes.

2 Overview of Relevant Technologies

2.1 Application of Classic LLM in Push Generation, Pushgen Framework

Pushgen framework uses controllable category prompts to constrain the generation of LLM, and can generate multiple different candidate documents in the same push scenario. Moreover, it introduces a reward model to sort the candidate documents, which can effectively make the output documents conform to the click through rate orientation and the output content specification of the platform. This framework has been deployed in fast hand, providing push for thousands of fast hand users [2].

This is the first verification of LLM in generating recommended documents on an industrial scale. The quality of the generated documents is equivalent to that of manual writing. The ranking mechanism of reward model in the framework effectively solves the uncontrollable problem of generating copy, and ensures the security of articles. At the same time, this framework also supports multi version generation, and the same push can also generate differentiated copies for different users.

In addition, the systematic comparison of different prompt strategies on the quality of notification text also shows that the optimized automatic prompt performs better in grammatical accuracy and consistency, but the manual prompt still has advantages in the diversity of text [5]. Pushgen framework also has its limitations. First, it has high dependence on manually designed controllable category prompts, and the quality of prompts will directly affect the quality of output documents. Second, the reward model also needs to be trained based on existing click data, and its adaptability to the cold start problem of new content cannot be verified. Finally, the cost of large-scale reasoning is high, and there is no specific optimization scheme.

2.2 BPE Subword Level Embedding Initialization Helps Solve the Representation of Cold Start Items

Split the text description of the new item into Byte Pair Encoding (BPE) sub word units, and extract the LLM embedded representation of each token [4]. Aggregate subwords are embedded to generate item level dense semantic vector, which is used as the initial value of new item embedding in the recommendation model to replace the traditional random initialization.

This method has the following advantages: first, it has a strong semantic priori. After the new item goes online, it has a rich semantic vector representation. Unlike the traditional method, it lacks new content data. Second, the content of BPE subwords is more accurate, the granularity is more fine, and the compatibility of the architecture is good, and the change cost is low.

Similarly, such a method also has the problems that it depends on the quality of the text information of the item, can not capture the dynamic changes of the interest distribution of the item, and may lose the interactive semantics between subwords when aggregating subwords.

Similar to this idea, mi4rec further proposed a meta item embedding method, which uses a small number of learnable tokens to dynamically generate a new item representation combined with the content of the item, and has made significant improvements in e-commerce and short video data sets, which can be used as a complementary option for BPE embedding scheme [6].

2.3 Joint Evaluation of Zero Sample, Context Learning and Fine Tuning

This article systematically compares three LLM usage strategies under the unified retrieve and recommend framework. One is zero sample prompt, which only gives LLM task instructions and descriptions of users and items, rather than traditional examples. The second is context learning, which provides a small number of input-output counterexamples as the prefix of prompts. The third is joint fine-tuning, which uses the mixed data of cold start and hot start to make light-weight fine-tuning of LLM, so that a single model can handle different scenarios at the same time [5]. This paper gives a comparative conclusion from the two dimensions of accuracy and calculation cost.

This is the first time to fairly compare the advantages and disadvantages of the three strategies under the conditions of cold start and hot start in the same framework. The conclusion has direct engineering guiding significance, and the joint fine-tuning simplifies the system architecture and reduces the maintenance cost.

However, at the same time, joint fine-tuning requires a certain amount of cold start data samples, and the push effect on absolute zero sample users needs to be evaluated. In addition, the experiment is mainly in the offline environment, lacking the test in the real online environment.

On this basis, section 3 will further analyze the application of these technologies in specific push

scenarios according to the two dimensions of new users and new content.

3 Application Analysis

Based on the three core technologies introduced in Section 2, LLM push generation (pushgen), BPE sub word level semantic embedding and LLM strategy comparison framework [2][7][8]. This section specifically shows how these technologies can be applied in the actual cold start push scenario from the two dimensions of new users and new content, and complements the methods of image generation, context learning and meta learning, forming a complete application link [4][9][10].

3.1 Cold Start of New Users

This article will show in 3.1 how LLM can complete the complete link from understanding who the user is to deciding what to push and finally generating what files in the absence of new user data. Figure 1 illustrates the complete AI-driven marketing chain: from understanding the user, to deciding what to push, and finally generating personalized copy.

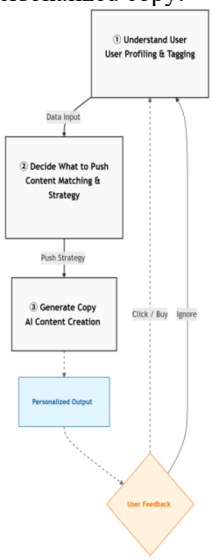

Figure 1. The AI-Powered Marketing Full-Chain: From User Understanding to Content Generation (Picture credit: Original)

3.1.1 *Implicit Portrait and Push Generation Based on Weak Signal*

Turn weak signals such as equipment model, installation period and geographical location into natural language prompts, and use context learning to make LLM infer user interest profile [10]. After that, call pushgen to generate multiple candidates and reward sorting to generate the first personalized push file that adapts to the profile, and personalize the original rigid push file [2].

3.1.2 *Actively Acquire Preferences and Use Intention Probes*

For new users without any behavior data, take the initiative to add some common questions in the welcome copy, such as "do people care more about sports or technology". LLM can analyze data such as users' clicks or replies in real time, update user profiles and trigger the second accurate push. This method gets rid of the shackles of traditional methods that can only wait for data, and uses the way of active dialogue to obtain the information of new users, but at the same time, it should also be noted that the questions should be concise and accurate.

3.1.3 *Use Meta Learning to Quickly Adapt to New Users*

After a new user generates 1-3 clicks, it uses the meta learning framework (maml/reply) to fine tune the prompt parameters of LLM in milliseconds [4]. This method is deeper than context fine-tuning and less than full-scale fine-tuning. Combined with the advantages of both, it can generate accurate personalized push for users after obtaining a small amount of data, and the delay is acceptable.

3.2 New Content Cold Start

3.2.1 *BPE Semantic Embedding for Instant Recall*

When a new item goes online, it uses LLM to encode its title/description at the BPE sub word level, and generates a dense semantic vector as the initial embedding [7]. This vector will be directly connected to the recall channel, so that new items can also be hit by the interest vector in the case of lack of data, which solves the "cold wait" problem of the traditional method.

3.2.2 *Adaptive Matching of Copy and Portrait*

For the same new item, use pushgen to generate multiple versions of copywriting emphasizing different selling points [2]. Combined with the LLM of new users to generate portraits, the system adaptively selects the most matching copy version to push and improve the touch efficiency of a single new product [9].

3.2.3 *Retrieval Enhancement Generation to Ensure the Accuracy of Facts*

When a new item lacks information description, retrieve the item details or external knowledge sources, embed such information into LLM for constraint, and access small model review if necessary. This method can effectively alleviate the risk of new content fabrication, but will slightly increase the delay.

3.3 Application Summary

The more accurate the user portrait, the higher the matching degree of new content push. Rich multi

version generation in turn improves the first retention of new users. New users and new content are deeply coupled. On this basis, section 4 will focus on the hallucinations, delays, evaluation and privacy challenges faced by these methods.

4 Discussion

4.1 Core Challenges

This paper mainly discusses the four challenges of current methods

First, the illusion still exists. LLM may push non-existent fabricated information. Although the retrieval enhancement and verification model can alleviate it, there is no way to cure it.

The second is the problem of reasoning delay. Multi version copy generation, retrieval enhancement and meta learning adaptation have doubled the number of LLM calls, which is prone to high delay under the conditions of real-time push and large-scale user push.

Third, there is a lack of evaluation benchmark, the performance of the test varies greatly between online and offline environments, the traditional CTR indicators cannot be used in the cold start scenario, and there is a lack of professional multidimensional evaluation framework.

Fourth, privacy and user trust. Collecting user information will inevitably lead to user privacy related issues. Whether users are willing to provide their own environmental data for weak signal inference and intention probe is questionable.

4.2 Outlook

In the future, with the development of technology, first of all, the paper hopes to deploy lightweight LLM to the device side, complete the portrait and generation only using local data, and solve the problems of delay, cost and privacy synchronously. Second, the paper hopes to use LLM to generate diversified virtual users, batch test the cold start strategy, and establish a low-cost, reproducible offline evaluation system. Finally, uplift modeling is introduced to evaluate the incremental causal effect of push, so that LLM can give priority to users who really need push guidance and avoid invalid interruption.

5 Conclusion

Starting from the cold start dilemma of "new user zero behavior" and "new content zero interaction" in push notifications, this paper systematically combs the content generation and personalization methods of the large language model in this scenario. The core judgment of the full text can be summed up in three points: first, the fundamental value of LLM is to fill the semantic blind spot of traditional methods under the condition of zero data, and make the cold start push from popular to zero sample personalization; Second, new users and new content are two asymmetrical cold start problems. The former is the inference task of "guess who" and the latter is the representation task of "describe what". The technical solutions should be designed separately; Third, under the current millisecond delay and 100 million user size constraints, the hybrid architecture is a realistic landing path that takes into account semantic quality and computing cost, and the pure LLM scheme is more suitable for offline scenarios. Looking forward to the future, local reasoning of end-to-side small models, off-line evaluation driven by generative user simulation, and causal inference to measure incremental value are three key breakthroughs that support each other. From a longer-term perspective, cold start push is moving from one-way prediction to two-way dialogue. LLM will go beyond the role of copywriter and become a personalized agent to understand users' intentions and deliver platform value. This is not only a technical proposition, but also a trust and ethical proposition that the next generation of push system must respond to.

References

1. Liu, C., Jiang, D., Cai, Y., & Li, H. (2026). A review of deep learning and large language models for cold start problem in recommender systems. WIREs Data Mining and Knowledge Discovery, 16(1), e70068. https://doi.org/10.1002/widm.70068
2. Bie, S., Cao, J., Luo, Z., et al. (2026). PushGen: Push notifications generation with LLM. In Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM ’26). ACM. https://doi.org/10.1145/3773966.3779373
3. Sanner, S., et al. (2026). Large language models are competitive near cold-start recommenders for language-based preferences. arXiv preprint / OpenReview.
4. Zhao, Y., Shen, H., Li, D., et al. (2025). Meta-learning for cold-start personalization in prompt-tuned LLMs. In Proceedings of the IEEE Conference on Big Data and Smart Engineering (CBASE 2025) (pp. 462–467). IEEE. https://arxiv.org/abs/2507.16672
5. Taşköprü, H.,et al. (2025). Notification text generation with large language models. In Proceedings of the IEEE Signal Processing and Communications Applications Conference (SIU 2025). Istanbul, Turkey.
6. Zheng, Z., Zhu, Y., Liu, H., et al. (2025). MI4Rec: Pretrained language model based cold-start recommendation with meta-item embeddings. In Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM ’25). ACM.
7. Zhao, Y., et al. (2025). Efficient cold-start recommendation via BPE token-level embedding initialization with LLM. arXiv preprint arXiv:2509.13179.

8. Halimeh, H., Freese, F., & Müller, O. (2025). LLMs for warm and cold next-item recommendation: A comparative study across zero-shot prompting, in-context learning and fine-tuning. In Proceedings of the International Conference on Information Systems Development (ISD). https://doi.org/10.62036/isd.2025.68
9. Wongso, W., et al. (2025). GenUP: Generative user profilers as in-context learners for next POI recommender systems. arXiv preprint arXiv:2410.20643v2.
10. Liang, X., Nguyen, V., Le, V., et al. (2025). In-context learning for addressing user cold-start in sequential movie recommenders. In Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys ’25). ACM.