Rebuttal of *Online Item Cold-Start Recommendation with Popularity-Aware Meta-Learning*, KDD 2025, Submission 195

Anonymous Author(s)

REVIEWER HVOV

Thank you for the thoughtful review of our work! Please allow us to address your concerns and answer the questions.

W1&Q3: Unclear definition of cold-start items.

The segmentation about cold-start items is an important part of our proposed meta-learning approach and involves the definition of the different tasks of meta-learning. As described in Section 5.1.2, cold-start items are defined by a threshold of low popularity. **This definition only takes into account the popularity of the item in the current period, so cold-start items may become popular in future periods.** This property of the same item entering different tasks in different periods is also an important theoretical basis for our cold-start enhancer.

W2: Uncommon InfoNCE loss and negative sampling.

The loss function form of InfoNCE is frequently used in both online and offline systems[1-3], due to its focus on correctly ordering users and items within a batch as opposed to more accurately predicting absolute scores. **It is therefore more suited on retrieving users and items in large systems, and we attempt to replicate this in our offline experiments.**

About negative sampling, since **we need to compute the recall and NDCG evaluation metrics about the items, it is necessary to collect negative samples of users for the items.** The reason for calculating the metrics for recommending users to items will be explained in W5.

W3: About fixed task segmentation.

Fixed task segmentation is one of the innovations of our proposed PAM. As stated in the Abstract and Introduction, **we simply adopt a meta-learning scheme to be used for sharing information while ensuring task specificity, and do not generate unique parameters for each cold-start task as in traditional meta-learning**. In PAM, cold-start items make up the task as a whole, paralleled by tasks made up of popular items. This approach allows the meta-learning process to differentiate between cold-start and popular item recommendations, capture properties common to cold-start items, and avoid the real-time computational overhead of new cold-start items.

W4: Necessity of hidden-layer meta-updates.

Since the same item is only likely to be present in the same task at the same moment in time, and there is no inter-task sharing, it does not make sense to meta-update the embedding parameters. For the hidden layer parameters, as analysed in Section 5.4, **since behaviour-based and content-based embedding have different importance in the recommendation of cold-start and popular items, the hidden layer parameters need to be personalised to the tasks in order to leverage these embedding differently.**

W5: Metrics of items.

In the PAM setup, different items go into different tasks. **In the evaluation, we want to check about the effectiveness of the recommended parameters in a single task, so we calculate metrics on items.** In the actual online system operation, for cold-start items, the system will retrieve users for the items and recommend them to the users for the purpose of cold-start.

W6&Q2: Cold start settings.

In our evaluation, we were consistent for all methods. We selected cold-start items and evaluated metrics for recalled users, and we do the same for sequential recommendation schemes such as ASMG, SML, and IMSR.

Our proposed PAM is a model-agnostic approach, so it can be overlaid on all types of methods. For example, we selected DIN[4] as the base model (Embedding&MLP in the article). The results are summarized by Table 1:

Method	Recall@5	NDCG@5
PF	0.6474	0.4876
PAM-M	0.8629	0.7208

Table 1: PAM's performance compared with PF with base model DIN.

As can be seen from the results, PAM still has a similar magnitude of improvement compared to the baseline under different base models.

W7&Q1: Detail of online A/B tests.

We implement PAM on one of the world's largest short-video platforms. We carried the online A/B testing in our short-video streaming scenario from Oct. 4-7, 2023, with hundreds of millions of users per day. For company privacy, we cannot report the implementation details such as the size of online data and the real performance of the original online models, thus we only report the improvement of metrics.

[1] Li et al. Contextual Distillation Model for Diversified Recommendation, KDD '24

[2] Cai et al. LightGCL: Simple Yet Effective Graph Contrastive Learning for Recommendation, ICLR '23

[3] Wu et al. Self-supervised Graph Learning for Recommendation, SIGIR '21

[4] Zhou et al. Deep Interest Network for Click-Through Rate Prediction. KDD '18.

REVIEWER KZRJ

Thank you for the thoughtful review of our work! Please allow us to address your concerns and answer the questions.

W: Online deployment of PAM.

As introduced in Section 4.2.2, **the online deployment of PAM does not require any additional real-time computational overhead.** Alg. 1 details the steps of the run. Firstly, at the current period, the arriving data is divided into different tasks for training according to their popularity, and **the task personalization parameters obtained in this step are stored.** And in the next period, the cold-start items **only need to obtain the parameters of the cold-start tasks that were computed in the previous moment and are directly inferred.**

In the training phase, since different items will not overlap between tasks, the computational overhead is the same as that of a common two-tower structure. If the meta-learning inner loop use multi-step scheme, the computational overhead will only grow by a few times accordingly, and there is no extra overhead due to the existence of the outer loop. **Compared to a normal dual-tower, PAM requires only a few times more space for storing the parameters of different tasks, and if only cold-start items are recommended in online serving, it is also feasible to store only the cold-start parameters, so there is no additional storage overhead.**

Q: Detail of online A/B tests.

We implement PAM on one of the world's largest short-video platforms. We carried the online A/B testing in our short-video streaming scenario from Oct. 4-7, 2023, with hundreds of millions of users per day. For company privacy, we cannot report the implementation details such as the size of online data and the real performance of the original online models, thus we only report the improvement of metrics.

REVIEWER GMJ8

Thank you for the thoughtful review of our work! Please allow us to address your concerns and answer the questions.

W1: Efficiency evaluation of PAM.

As introduced in Section 4.2.2, **the online deployment of PAM does not require any additional real-time computational overhead.** Alg. 1 details the steps of the run. Firstly, at the current period, the arriving data is divided into different tasks for training according to their popularity, and **the task personalization parameters obtained in this step are stored.** And in the next period, the cold-start items **only need to obtain the parameters of the cold-start tasks that were computed in the previous moment and are directly inferred.**

However, for the traditional meta-learning approach, **when a new item arrives, the recommended parameters for that item require a step of fine-tuning of the meta-parameters in order to be obtained** (as is also the case for FORM), since each item is treated as a separate task, and cannot be obtained directly in the previous task as in PAM. This step of fine-tuning inevitably brings time-consumption in real-time inferring, whereas PAM does not have any time-consumption in this step. Therefore, we consider that a numerical evaluation of the time-consumption in the online service is not necessary.

W2&3: Presentation and symbol problems.

Thank you for pointing this out, the symbol issue you described is the result of an oversight in our work and we will fix the issue you mentioned in camera-ready version.

W4: Optimization of figures.

Thank you for pointing this out, we have found the issue you raised about Fig. 2, the picture is indeed not clear enough. We have modified the picture so that it more accurately depicts the model and does not cause confusion. For the revised image with the manuscript, please see refer to the files.

Q1: Ablation study about different embeddings.

We summarise a set of experiments in which different parts of the embedding are retained. The results are shown in Tab. 2.

In the table, PAM-M-Beh represents the model that only retain ID and historical behavior embeddings, PAM-M-Con represents the model that retain ID and content-based embeddings. From the results, we can find that retaining only part of the embedding leads to worse recommendation. Meanwhile, content-based embedding is more important in the recommendation of cold-start items, and retaining content-based embedding is more effective than retaining behaviour-based embedding.

We further performed experiments similar to the breakdown analysis for PAM-M-Beh, and the results are shown in Fig. 1. For cold-start items and popular items, their ID embedding and behavioural sequence embedding still play different roles. In cold-start items, the ID embedding has almost no information, while the few users it interacts with may contain more information than ID. In contrast, in popular items, their ID embedding has been well learnt and plays a more important role. The visualisation results confirm this and illustrate the effectiveness of PAM's meta-learning approach in different task specifications.

Table 2: The reported top-K evaluation metric results on MovieLens dataset.

	R@5	R@10	R@20	N@5	N@10	N@20
PAM-M	0.3846	0.4733	0.5727	0.3040	0.3327	0.3578
PAM-M-Beh	0.2648	0.3410	0.4317	0.1987	0.2233	0.2462
PAM-M-Con	0.3544	0.4435	0.5392	0.2767	0.3054	0.3296

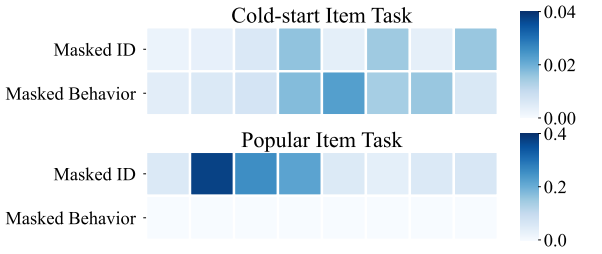

Figure 1: Squared errors of top representations of cold-start items and popular items before and after masking for different types of embedding inputs (ID and historical behavior).

Q2: Adding more baseline.

To the best of our knowledge, we are the first to present an online cold-start approach that does not require real-time fine-tuning. For FORM, its treats each individual user as a task and computes personalised parameters for each user as they arrive with fine-tuning. This approach is unacceptably time-consuming in large online systems and can only be deployed in online systems with day-level updates, so we do not compare FORM.

Your suggestion is very important to us and we are in the process of reproducing the code of FORM. Due to the time constraints of rebuttal and the fact that FORM is not open source, it will still take some time to reproduce it, so we will add our comparison evaluations in the camera-ready version.

REVIEWER KPSE

Thank you for the thoughtful review of our work! Please allow us to address your concerns and answer the questions.

W1: Write up issues.

Thanks for pointing this out, we'll be checking the article in detail and fixing the grammatical issues you mentioned in the camera-ready version.

W2: About sequential and online recommendation.

The writing here is not a clerical error. In our view, sequence recommendation and online recommendation are different subproblems of recommendation domains. **Sequential recommendation focuses on the interaction history of a user or an item and extracts information from it such as user interests that are beneficial for future recommendations**. This process does not need to be online, offline recommendations can also be made sequentially. **Online recommendation, on the other hand, has different requirements than sequential recommendation, focusing more on the linearity and timeliness of the recommendation, and has requirements on the computational overhead of the system**, but it does not have to use the history of user-item interaction information. In our PAM, we use historical interaction sequences only as part of behaviour-based embedding, not for the purpose of sequence recommendation.