A Experimental Setup

Data Splits. Following the setting in our KDDCUP competition[3], the *Amazon-M2* dataset contains three splits: training, phase-1 test and phase-2 test. For the purpose of model training and selection, we further split the original training set into 90% sessions for development (used for training), and 10% sessions for validation. Note that the numbers in Tables 3, 4, and 5 of the main content are for validation performance. Without specific mention, the test set mentioned in the main content indicates the phase-1 test. Due to the page limitation of main content, we defer the performances on phase-2 test set to the appendix.

Hyperparameter Settings. The hyper-parameters of all the models are tuned based on the performance on validation set. For Task 1 and Task 2, we follow the suggested hyper-parameter range to search for the optimal settings. By default, we only use the product ID to train the models since most of the popular session-based recommendation baselines are ID-based methods. We leave the exploration of other rich attributes such as price, brand, and description as future work. Specifically, the search ranges for different models are outlined below:

- GRU4REC++ : learning_rate [0.01,0.001,0.0001], dropout_prob: [0.0,0.1,0.2,0.3,0.4,0.5], num_layers: [1,2,3], hidden_size: [128].
- NARM: learning_rate: [0.01,0.001,0.0001], hidden_size: [128], n_layers: [1,2], dropout_probs: ['[0.25,0.5]','[0.2,0.2]','[0.1,0.2]'].
- STAMP: learning_rate: [0.01,0.001,0.0001]
- SRGNN: learning_rate: [0.01,0.001,0.0001], step: [1, 2].
- CORE: learning_rate: [0.001, 0.0001], n_layers: [1, 2], hidden_dropout_prob: [0.2, 0.5], attn_dropout_prob: [0.2, 0.5].
- GRU4RECF: learning_rate: [0.01,0.001,0.0001], num_layers: [1, 2].
- SRGNNF: learning_rate: [0.01,0.001,0.0001], step: [1, 2].

For Task 3, we tune the following hyperparameters in mT5: weight_decay in the range of {0, 1e-8}, learning_rate in the range of {2e-5, 2e-4}, num_beams in the range of {1, 5}. Additionally, we set the training batch size to 12, and the number of training epochs to 10.

Hardware and Software Configurations. We perform experiments on one server with 8 NVIDIA RTX A6000 (48 GB) and 128 AMD EPYC 7513 32-Core Processor @ 3.4 GHZ. The operating system is Ubuntu 20.04.1.

B More Dataset Details

B.1 Dataset Collection

The *Amazon-M2* dataset is a collection of anonymous user session data and product data from the Amazon platform. Each session represents a list of products that a user interacted with during a 30-minute active window. Note that the product list in each session is arranged in chronological order, with each product represented by its ASIN number. Users can search for Amazon products using their ASIN numbers[4] and obtain the corresponding product attributes. We include the following product attributes:

- ASIN (id): ASIN stands for Amazon Standard Identification Number. It is a unique identifier assigned to each product listed on Amazon's marketplace, allowing for easy identification and tracking.
- Locale: Locale refers to the specific geographical or regional settings and preferences that determine how information is presented to users on Amazon's platform.
- Title: The Title attribute represents the name or title given to a product, book, or creative work. It provides a concise and identifiable name that customers can use to search for or refer to the item.

[3] https://www.aicrowd.com/challenges/amazon-kdd-cup-23-multilingual-recommendation-challenge

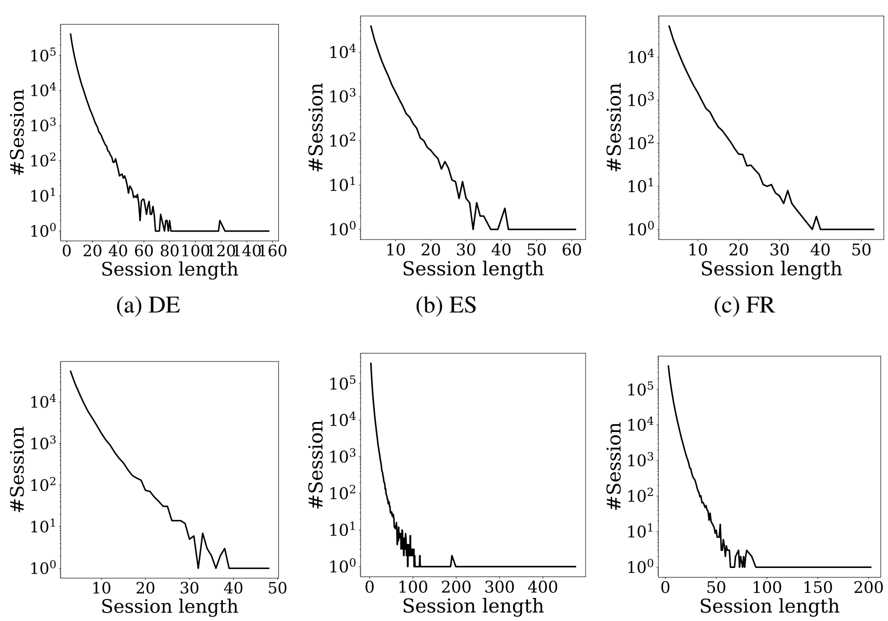
[4] For instance, if the product is available in the US, users can access the product by using the following link: https://www.amazon.com/dp/ASIN_Number

- Brand: The Brand attribute represents the manufacturer or company that produces the product. It provides information about the brand reputation and can influence a customer's purchasing decision based on brand loyalty or recognition.
- Size: Size indicates the dimensions or physical size of the product. It is useful for customers who need to ensure that the item will fit their specific requirements or space constraints.
- Model: The Model attribute refers to a specific model or version of a product. It helps differentiate between different variations or versions of the same product from the same brand.
- Material Type: Material Type indicates the composition or main material used in the construction of the product. It provides information about the product's primary material, such as metal, plastic, wood, etc.
- Color Text: Color Text describes the color or color variation of the product. It provides information about the product's appearance and helps customers choose items that match their color preferences.
- Author: Author refers to the individual or individuals who have written a book or authored written content. It helps customers identify the creator of the work and plays a significant role in book purchasing decisions.
- Bullet Description (desc): Bullet Description is a concise and brief description of the product's key features, benefits, or selling points. It highlights the most important information about the item in a clear and easily scannable format.

The dataset spans a period of 3 weeks, with the first 2 weeks designated as the training set and the remaining week as the test set. To enhance evaluation, the test set is further randomly divided into two equal subsets, i.e., phase-1 test and phase-2 test.

B.2 Additional data analysis

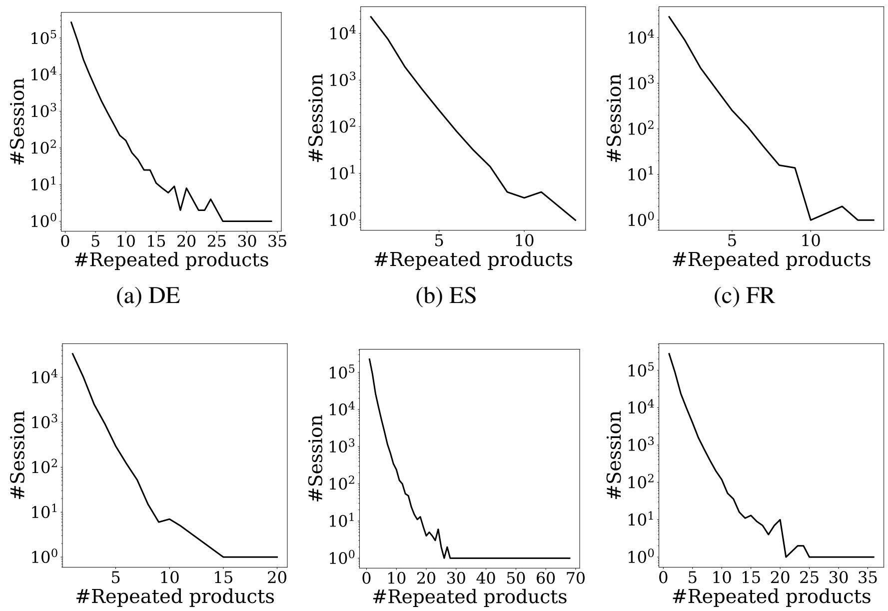
The additional analysis of the session length and the repeat pattern on each individual locale can be found in Figure 3 and 4, respectively. We can still observe evident long-tail distributions for both product frequency and repeat pattern across the six locales, which is consistent with the observation that we made on the whole dataset in Section 3.



(a) DE (b) ES (c) FR

(d) IT (e) JP (f) UK

Figure 3: Session length w.r.t. locales where the x-axis corresponds to the session length (the number of items in a session), the y-axis indicates the number of sessions with the corresponding session length. A clear long-tail phenomenon can be found, where only a few sessions show a session length of more than 100.



(a) DE (b) ES (c) FR

(d) IT (e) JP (f) UK

Figure 4: The number of repeat items w.r.t. locales where the x-axis corresponds to the number of repeat items, the y-axis indicates to the number of session with the corresponding number of repeat items. Notably, we exclude those sessions with no repeat patterns. A clear long-tail phenomenon can be found, where only a few sessions show many repeat items.

B.3 License

The *Amazon-M2* dataset can be freely downloaded at https://www.aicrowd.com/challenges/amazon-kdd-cup-23-multilingual-recommendation-challenge/problems/task-1-next-product-recommendation/dataset_files and used under the license of Apache 2.0. The authors agree to bear all responsibility in case of violation of rights, etc.

B.4 Extended Discussion

Item Cold-Start Problem. The item cold-start problem [66, 34] is a well-known challenge in recommender systems, arising when a new item is introduced into the system, and there is insufficient data available to provide accurate recommendations. However, our dataset provides rich items attributes including detailed textual descriptions, which offers the potential to obtain excellent semantic embeddings for newly added items, even in the absence of user interactions. This allows for the development of a more effective recommender system that places greater emphasis on the semantic information of the items, rather than solely relying on the user's past interactions. Therefore, by leveraging this dataset, we can overcome the cold-start problem and deliver better diverse recommendations, enhancing the user experience.

Data Imputation. Research on deep learning requires large amounts of complete data, but obtaining such data is almost impossible in the real world due to various reasons such as damages to devices, data collection failures, and lost records. Data imputation [67] is a technique used to fill in missing values in the data, which is crucial for data analysis and model development. Our dataset provides ample opportunities for data imputation, as it contains entities with various attributes. By exploring different imputation methods and evaluating their performance on our dataset, we can identify the most effective approach for our specific needs.

Table 7: Experimental results on Task 1 phase-1 test set.

	MRR@100				Recall@100			
	UK	DE	JP	Overall	UK	DE	JP	Overall
Popularity	0.2723	0.2746	0.3196	0.2875	0.4940	0.5261	0.5652	0.5262
GRU4Rec++	0.2094	0.2082	0.2527	0.2222	0.4856	0.5192	0.5416	0.5137
NARM	0.2235	0.2233	0.2705	0.2378	0.5220	0.5594	0.5814	0.5524
STAMP	0.2398	0.2398	0.2888	0.2547	0.4265	0.4538	0.4867	0.4538
SRGNN	0.2240	0.2211	0.2670	0.2361	0.4986	0.5311	0.5540	0.5262
CORE	0.1777	0.1797	0.2103	0.1882	0.6513	0.6927	0.7009	0.6801

Table 8: Experimental results on Task 1 phase-2 test set.

	MRR@100				Recall@100			
	UK	DE	JP	Overall	UK	DE	JP	Overall
Popularity	0.2711	0.2754	0.3205	0.2875	0.4937	0.5283	0.5660	0.5271
GRU4Rec++	0.2081	0.2097	0.2533	0.2224	0.4843	0.5220	0.5420	0.5143
NARM	0.2219	0.2235	0.2720	0.2377	0.5209	0.5624	0.5786	0.5522
STAMP	0.2387	0.2402	0.2894	0.2546	0.4234	0.4585	0.4864	0.4541
SRGNN	0.2224	0.2224	0.2695	0.2367	0.4974	0.5336	0.5529	0.5262
CORE	0.1755	0.1807	0.2111	0.1880	0.6518	0.6966	0.7002	0.6813

C More Experimental Results

C.1 Task 1. Next-product Recommendation

In this subsection, we provide the model performance comparison on the phase-1 test and phase-2 test in Table 7 and Table 8, respectively. We can have similar observations as we made in Section 4.1: the popularity heuristic generally outperforms the deep models with respect to both MRR and Recall, with the only exception that CORE achieves better performance on Recall. This suggests that the popularity heuristic is a strong baseline and the challenging *Amazon-M2* dataset requires new recommendation strategies to handle. We believe that it is potentially helpful to design strategies that can effectively utilize the available product attributes.

C.2 Task 2. Next-product Recommendation with Domain Shifts

We report the mode performances on phase-1 test and phase-2 test in Table 9 and Table 10, respectively. Note that we omit the supervised training results since we have already identified that finetuning can significantly improve it. From the tables, we arrive at a similar observation as presented in Section 4.2 that the finetuned deep models generally outperform the popularity heuristic in Recall but underperform it in MRR. This illustrates that the deep models have the capability to retrieve a substantial number of pertinent products, but they fall short in appropriately ranking them. As a result, there is a need to enhance these deep models further in order to optimize their ranking efficacy.

C.3 Task 3. Next-product Title Prediction

We expand Table 6 to include the results on phase-2 test and the full results are shown in Table 11. From the table, we have the same observations as we made in Section 4.3: (1) Extending the session history length (K) does not contribute to a performance boost, and (2) The simple heuristic of Last Product Title outperforms all other baselines. It calls for tailored designs of language models for this challenging task.

Table 9: Experimental results on Task 2 phase-1 test.

		MRR@100				Recall@100			
	Methods	ES	FR	IT	Overall	ES	FR	IT	Overall
Heuristic	Popularity	0.2934	0.2968	0.2887	0.2927	0.5725	0.5825	0.5861	0.5816
Pretraining & finetuning	GRU4Rec++	0.2665	0.2829	0.2527	0.2669	0.6467	0.6612	0.6600	0.6573
	NARM	0.2707	0.2890	0.2608	0.2733	0.6556	0.6612	0.6685	0.6629
	STAMP	0.2757	0.2860	0.2653	0.2753	0.5254	0.5377	0.5371	0.5346
	SRGNN	0.2853	0.2979	0.2706	0.2840	0.6263	0.6505	0.6453	0.6427
	CORE	0.2058	0.2091	0.1984	0.2040	0.7457	0.7384	0.7545	0.7466

Table 10: Experimental results on Task 2 phase-2 test.

		MRR@100				Recall@100			
	Methods	ES	FR	IT	Overall	ES	FR	IT	Overall
Heuristic	Popularity	0.3017	0.3068	0.2902	0.2989	0.5818	0.5934	0.5826	0.5863
Pretraining & finetuning	GRU4Rec++	0.2648	0.2867	0.2569	0.2695	0.6473	0.6619	0.6600	0.6577
	NARM	0.2742	0.2938	0.2658	0.2779	0.6617	0.6624	0.6742	0.6670
	STAMP	0.2809	0.2922	0.2653	0.2787	0.5340	0.5400	0.5387	0.5381
	SRGNN	0.2878	0.3035	0.2701	0.2863	0.6359	0.6582	0.6491	0.6493
	CORE	0.2016	0.2138	0.1967	0.2040	0.7530	0.7387	0.7572	0.7495

Table 11: Full results of BLEU scores in Task 3.

	Validation	Phase-1 Test	Phase-2 Test
mT5-small, $K = 1$	0.2499	0.2265	0.2245
mT5-small, $K = 2$	0.2401	0.2176	0.2166
mT5-small, $K = 3$	0.2366	0.2142	0.2098
mT5-base, $K = 1$	0.2477	0.2251	0.2190
Last Product Title	0.2500	0.2677	0.2655

D Limitation & Broader Impact

The release of the *Amazon-M2* dataset brings several potential broader impacts and research opportunities in the field of session-based recommendation and language modeling. It provides the potential for research in the session recommendation domain to access the rich semantic attributes and knowledge from multiple locales, enabling better recommendation systems for diverse user populations.

While the *Amazon-M2* dataset offers significant research potential, it is crucial to consider the certain limitations associated with its use. Despite efforts have been made to include diverse user behaviors and preferences with multiple locales and languages, it may not capture the full linguistic and cultural diversity of all regions. Moreover, the dataset can be only collected within the Amazon platform, which may not fully capture the diversity of user behaviors in other domains or platforms, leading to a potential biased conclusion and may not hold true in different contexts.

We also carefully consider the broader impact from various perspectives such as fairness, security, and harm to people. No apparent risk is related to our work.