# BERT4Rec Ablation Study for Movie Recommendation

Ataklti Kidanemariam (s4590821) and Benard Adala Wanyande (s4581733)

Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands

**Abstract.** This study performs ablation tests on BERT4Rec hyperparameters for sequential movie recommendation using the MovieLens 1M dataset converted to implicit feedback. We evaluate the impact of masking ratio, hidden size, layers, sequence length, and attention heads on Recall@10 and NDCG@10, identifying optimal configurations for this task.

## 1 Methodology

We investigate BERT4Rec [3], a Transformer-based [4] model, for sequential recommendation.

**Dataset and Preprocessing:** The MovieLens 1M dataset [1] was used. Ratings $\geq 4$ became positive interactions (1); users with $< 5$ interactions were filtered. Chronological sequences were padded/truncated to a max length. Data was split by user (70% train, 15% validation, 15% test), with the last item used for evaluation.

**Model, Training, and Evaluation:** The BERT4Rec model employs Transformer blocks with multi-head self-attention. Training uses a masked item prediction objective (similar to MLM) with Adam optimizer [2], cross-entropy loss on masked items, and early stopping based on validation NDCG@10. Performance is measured by Recall@10 and NDCG@10.

## 2 Results: Ablation Studies

We systematically varied key hyperparameters. Table 1 summarizes the performance for each tested configuration, highlighting the best result within each parameter group in bold.

## 3 Discussion

The ablation results, summarized in Table 1, demonstrate that BERT4Rec's performance on MovieLens 1M is highly dependent on hyperparameter configuration. The preference for a larger hidden size (256) alongside a shallow depth (2 layers) is particularly noteworthy, deviating from common deep architectures

Table 1: BERT4Rec Hyperparameter Ablation Study Results.

| Hyperparameter | Value | Recall@10 | NDCG@10 |
|---|---|---|---|
| Masking Ratio | **0.15** | **0.4901** | **0.2705** |
| | 0.20 | 0.3238 | 0.1721 |
| | 0.25 | 0.4674 | 0.2590 |
| Hidden Size | 64 | 0.1471 | 0.0720 |
| | 128 | 0.1443 | 0.0709 |
| | **256** | **0.2863** | **0.1505** |
| Number of Layers | **2** | **0.4285** | **0.2333** |
| | 4 | 0.4096 | 0.2183 |
| | 6 | 0.2264 | 0.1142 |
| Sequence Length | 50 | 0.3269 | 0.1712 |
| | **100** | **0.3932** | **0.2119** |
| | 200 | 0.3699 | 0.1947 |
| Attention Heads | **2** | **0.4701** | **0.2588** |
| | 4 | 0.4670 | 0.2581 |
| | 8 | 0.4599 | 0.2533 |

in NLP. This suggests that for this sequential recommendation task, model capacity is crucial for capturing item nuances, but excessive depth might hinder performance, potentially due to optimization challenges or the nature of user sequences compared to denser linguistic data. The optimal sequence length of 100 reinforces the idea of balancing sufficient historical context against noise from older, less relevant interactions.

Furthermore, the optimal masking ratio of 0.15 indicates a delicate trade-off in the self-supervised training objective. While masking drives the learning process, masking too few items might provide insufficient training signal, whereas masking too many (as potentially seen with the dip at 0.20) could remove essential context needed for accurate predictions. The slight preference for fewer attention heads (2) suggests that for the chosen hidden dimension and the complexity of dependencies in MovieLens sequences, intricate subspace attention might not yield significant benefits and could even slightly degrade performance, possibly due to increased parameterization without corresponding gains in modeling power.

Interestingly, these results align with broader findings in recommendation systems literature, where shallower models with well-tuned feature spaces often outperform deeper, more complex networks when data sparsity or sequential patterns dominate. This suggests that domain-specific adjustments, such as limiting model depth and tuning the context window carefully, are crucial in real-world recommendation deployments. In the context of BERT4Rec, lightweight architectures may not only enhance performance but also significantly reduce

inference costs, an important consideration for large-scale deployment environments.

It is important to acknowledge that these findings are specific to the MovieLens 1M dataset and the exact experimental setup. The optimal hyperparameters might differ for other datasets with varying sparsity or sequence length distributions. Additionally, these ablation studies varied one parameter at a time, holding others constant (likely at some baseline or previously determined optimum). A more exhaustive search exploring the interplay between parameters (e.g., does the optimal number of layers change significantly with hidden size?) could potentially reveal further insights but is computationally expensive. Therefore, these results provide strong individual indicators but don't capture the full complexity of the hyperparameter space interactions.

## 4 Conclusion

This study systematically evaluated the impact of key hyperparameters on BERT4Rec for movie recommendation. Optimal settings identified include a masking ratio of 0.15, hidden size 256, 2 layers, sequence length 100, and 2 attention heads. These findings underscore the necessity of task-specific tuning for sequential recommendation models and provide a practical baseline for configuring BERT4Rec. Future work could investigate the interplay between these hyperparameters more deeply or explore adaptive masking strategies.

Beyond hyperparameter optimization, another promising direction for future research is the incorporation of dynamic user modeling, where the model adapts the importance of past interactions based on temporal recency or user-specific behavior patterns. Introducing such temporal biasing mechanisms could further refine the model's predictive ability, especially for users with evolving preferences. Additionally, lightweight ensembling methods, combining shallow Transformer architectures like BERT4Rec with complementary models (e.g., GRU4Rec), may offer robustness improvements without significant computational overhead.

## References

1. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM Transactions on Interactive Intelligent Systems (TiiS) **5**(4), 1–19 (2015)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
3. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., Jiang, P.: Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM international conference on information and knowledge management. pp. 1441–1450 (2019)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems 30. pp. 5998–6008 (2017)