Model Performance Diagnostic Questionnaire: (Generic)

Data Sanity Checks

1. Can model overfit a small subset of the data (e.g., 1 user, 5 samples)?
2. Are input features and labels correctly aligned (no off-by-one or leakage)?
3. Are the input dimensions and formats (e.g., float vs int, shape) consistent across all pipeline stages?
4. Are you handling padding, masks, or missing values properly?
5. Are the data distributions reasonable (e.g., class balance, sequence length, interaction frequency)?
6. Is the preprocessing consistent across training, validation, and test data?
7. Are the item/user IDs or tokens uniquely and consistently encoded?

Model Architecture and Implementation

8. Are the input, hidden, and output dimensions correctly configured across all layers?
9. Are the activations (e.g., ReLU, tanh, sigmoid) behaving as expected — not saturating or zeroed out?
10. Are embedding layers initialized correctly and allowed to update during training?
11. If applicable, is model correctly handling sequence ordering, recurrence, or attention flow?
12. Are residual or skip connections, if any, applied correctly?
13. Is output layer appropriate for the task (e.g., softmax for classification, sigmoid for multi-label)?

Training Setup and Optimization

14. Is loss function appropriate for the task and model output?
15. Are the gradients flowing through all layers (no silent gradient blockers)?
16. Are you using gradient clipping if the model is deep or unstable?
17. Is learning rate and optimizer (e.g., Adam, SGD) suitable for this architecture?
18. Are batch size and sequence length chosen to balance generalization and stability?
19. Are regularization techniques (dropout, weight decay) correctly implemented?

Logging, Visualization, and Analysis

20. Is training loss decreasing, or completely stagnant?
21. Are validation metrics improving, diverging, or showing signs of overfitting?
22. Are model weights or activations exploding, vanishing, or remaining unchanged?
23. Are you visualizing internal components (e.g., attention maps, hidden states, embeddings) for insight?

Isolation and Baseline Comparison

24. Can a simpler version of model (e.g., linear or shallow) learn on the same data?
25. Have you tried swapping in a known baseline (e.g., logistic regression, GRU, SASRec) to confirm the dataset and training pipeline are valid?

Model Performance Diagnostic Questionnaire: (Model Specific – xLSTM for example)

Data Sanity for XLSTM (1–7)

1. Can XLSTM overfit a single user's sequence (e.g., 5–10 items)?
2. Are input sequences sorted by time, not shuffled across steps or users?
3. Are target items correctly shifted (e.g., predicting x_{t+1} from x_t)?
4. Are item/user IDs properly indexed and embedded (e.g., 0 reserved for padding)?
5. Are sequences padded and masked consistently for variable-length input?
6. Do any sequences contain only padding or too-few interactions (<3 steps)?
7. Are timestamps, if used, processed and aligned with interactions (no misalignment)?

XLSTM-Specific Architecture (8–13)

8. Are LSTM hidden states and cell states being properly initialized (e.g., zeros or learned)?
9. Is sequence length preserved throughout time steps (no silent truncation or misalignment)?
10. Are you using gated extensions (like attention-LSTM or XLSTM enhancements) correctly wired?
11. Are LSTM gradients vanishing due to long sequences or poor initialization?
12. Are input embeddings, temporal embeddings, or other contextual features learned correctly?
13. Is final dense/prediction layer correctly interpreting the LSTM outputs (last state, mean pooling, attention over time)?

Training Setup for XLSTM (14–19)

14. Is the learning rate appropriate for the XLSTM's depth and capacity?
15. Are you using gradient clipping (e.g., at norm 1 or 5) to control exploding gradients?
16. Is the loss function (e.g., BPR, softmax cross-entropy) aligned with the output shape?
17. Are negative samples for contrastive/BPR losses sampled per user or globally (as needed)?
18. Is the batch size large enough to stabilize LSTM learning without starving the model?
19. Are you tracking the hidden state norms over time — do they vanish or explode?

Monitoring & Debugging XLSTM (20–23)

20. Is the loss decreasing over epochs, or flatlining?
21. Are metrics like Hit@K, NDCG@K improving on validation data?
22. Are hidden states or cell states becoming degenerate (e.g., near-zero or uniform)?
23. Are activation distributions (e.g., tanh, sigmoid outputs) healthy — or are gates always open/closed?

Baseline and Isolation Strategy (24–25)

24. Can a vanilla LSTM baseline (no XLSTM extensions) learn on the same data?
25. If you strip XLSTM to a basic RNN/LSTM, does learning start — helping localize the fault to XLSTM enhancements?