Towards the Better Ranking Consistency: A Multi-task Learning Framework for Early Stage Ads Ranking

Xuewei Wang

Meta

AdKDD 23

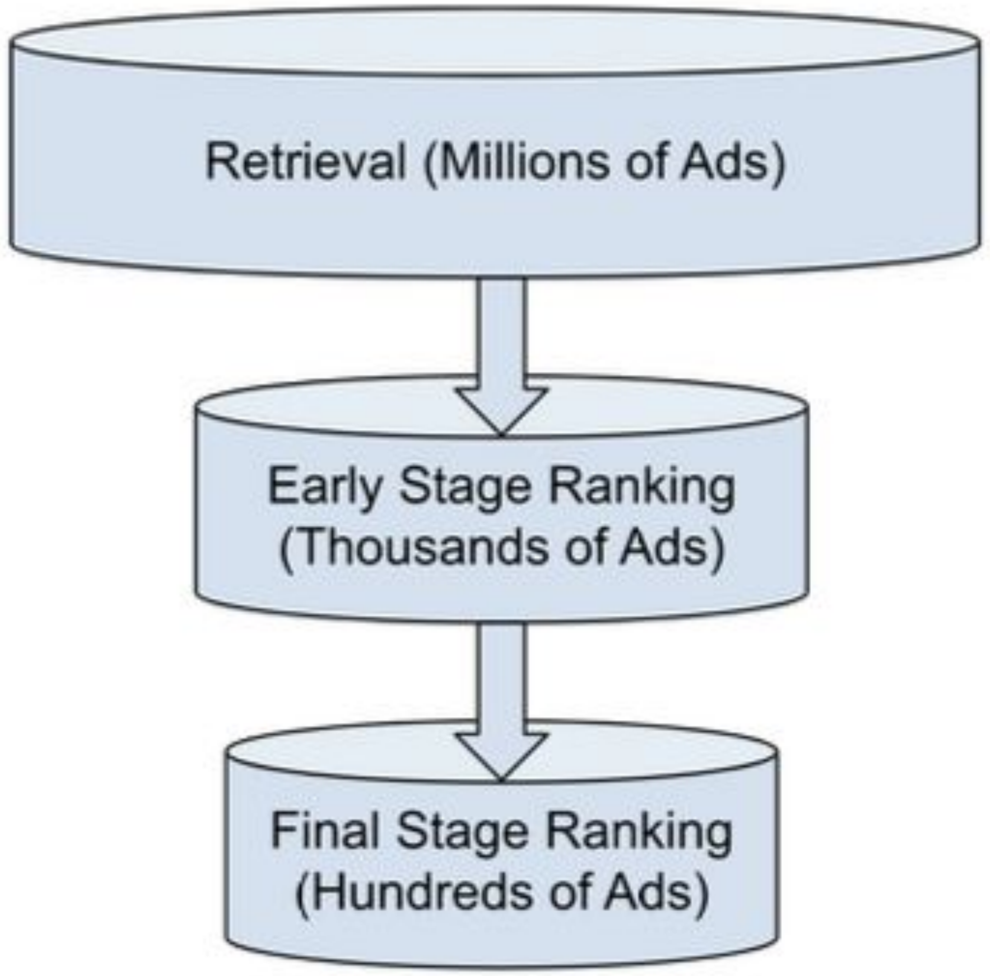
Work by: X Wang, Q Jin, S Huang, M Zhang, X Liu, Z Zhao, Y Chen, Z Zhang, J Yang, E Wen, S Chordia, W Chen, Q Huang

Meta

Ads Ranking system

- Multi-stage ranking system: Trade-off between capacity and efficiency
  - Early stage: Simplified model with latency constraint
  - Final stage: Large capacity model with good accuracy

- Multi-objective ranking system
  - Ad Auction depend on **Total Value**
    - Bid placed by an advertiser for that ad
    - ✅ Estimated action rates (e.g. CTR, CVR)
    - ✅ Ad Quality for user's ads experience
      - E.g. hide ads, report bad ads



Meta

Ranking Consistency

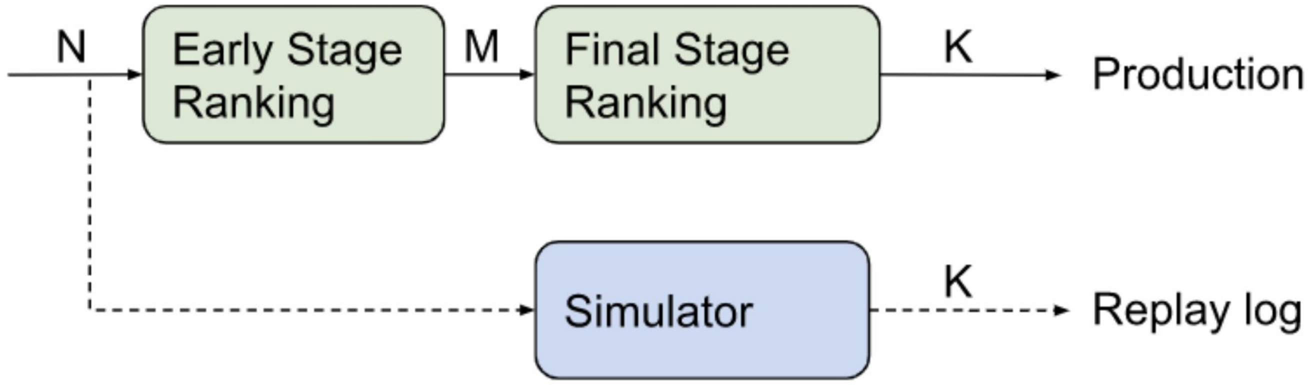
- Ideal Status
  - Early stage and final stage have same ranking orders for ads

- Ranking consistency issue:
  - Top ads in the final stage are ranked low in the early stage

- Gap between final stage and early stage:
  - Performance gap
  - Total value definition inconsistency
    - The early stage models' (i.e. ad quality models) development lags behind final stage models.
  - Selection bias

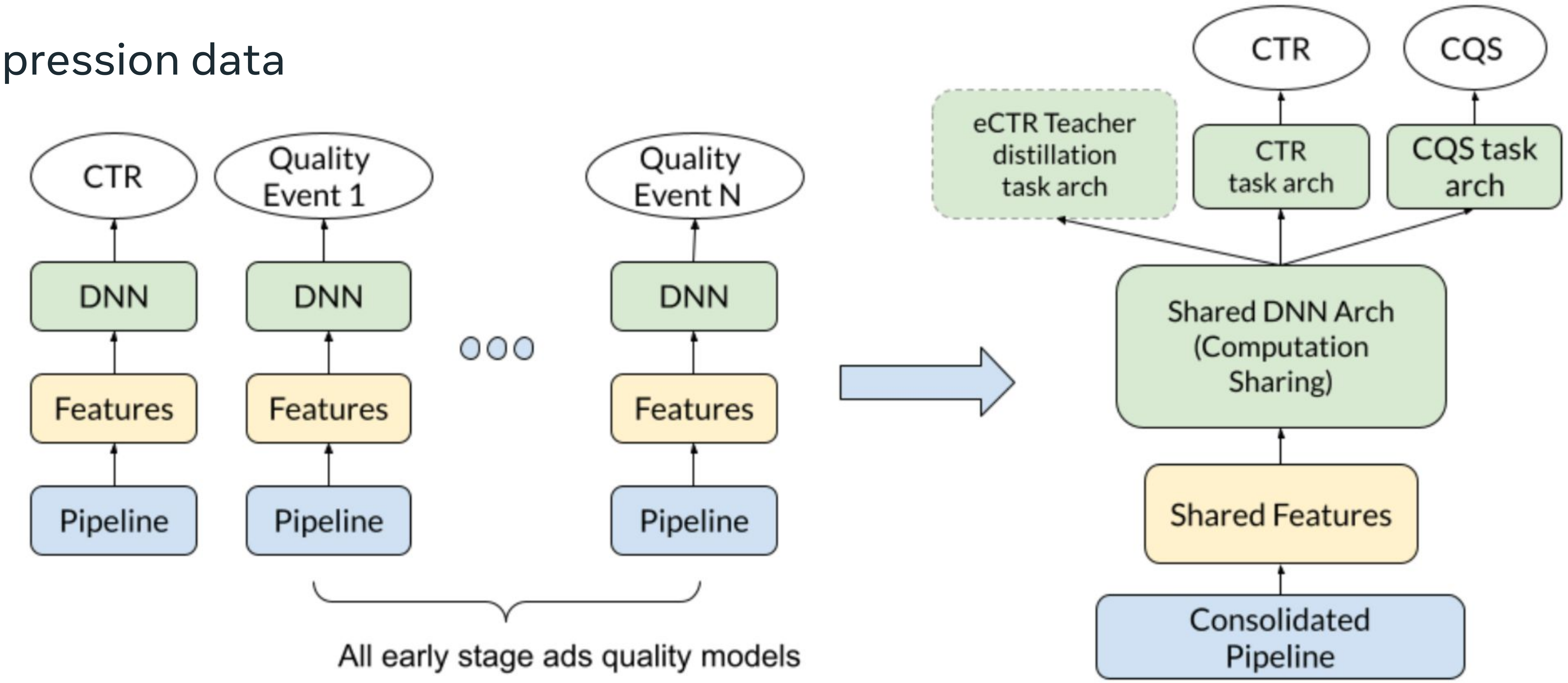
Ads recall for ranking consistency

- Why recall?
  - We only care about the top ad candidates for user impression, rather than the lower-ranked ones
- Challenge:
  - The accurate recall is difficult to compute with large candidates
- Solutions:
  - Offline simulated recall
    - We replay a small traffic with full ad requests in simulators, with relaxed timeouts between stages, to ensure that all ads from retrieval stages are ranked.



Multi-task learning for early stage ranking

- Ranking consistency improvement
  - Learn from final stage ads quality models
  - Learn from final stage CTR model
- Resource saving by model consolidation
  - Ads quality & CTR models need to predict on most ads
- Mitigation of selection bias
  - Data augmentation with non-impression data



Multi-task learning for early stage ranking

- New objective:
  - Consolidated Quality Score (CQS): Final stage total quality score

$$AdQuality = f(CQS)$$

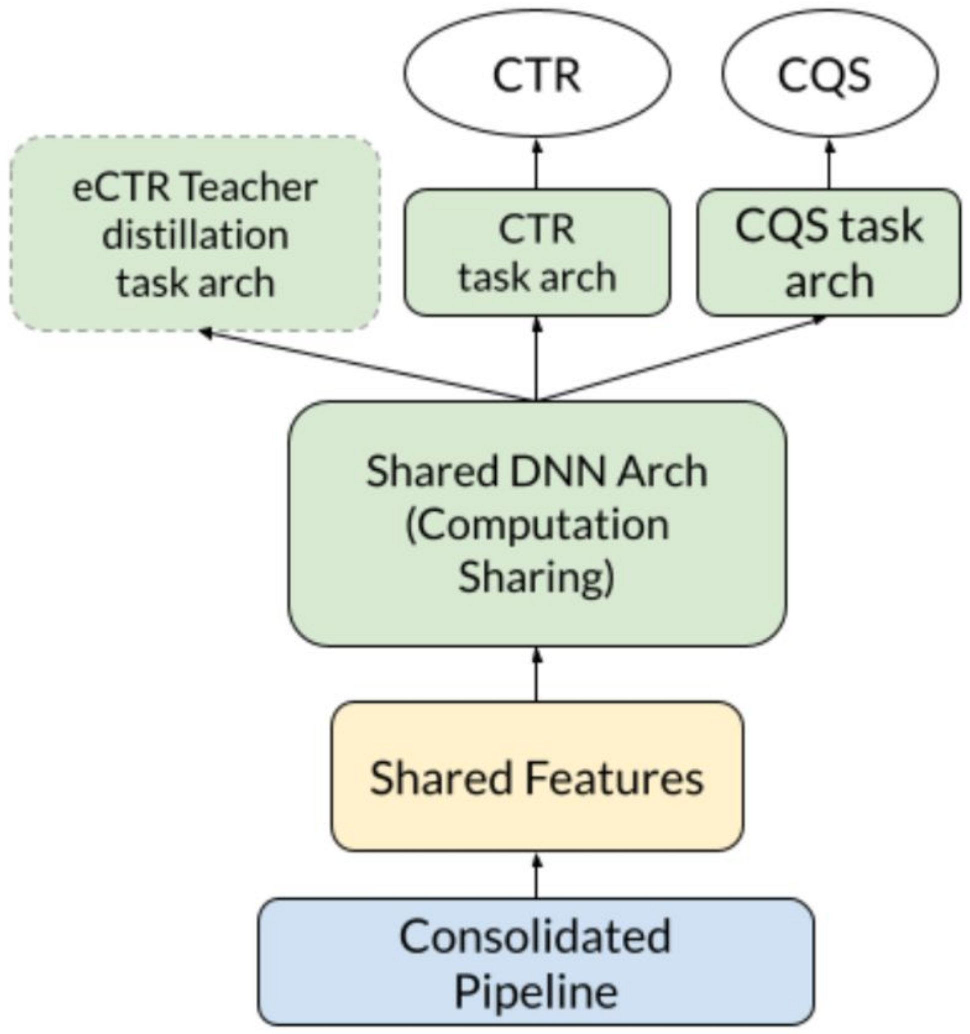
$$CQS = \sum_{i=1}^{N} scalar_i * pQualityEvent_i$$

$$L_{cqs} = \frac{1}{n} \sum_{i=1}^{n} (CQS_i - y_{cqs})^2,$$

- Final stage CTR teacher distillation

$$L_{teacher} = -[eCTR * \log(y_{ctr}) + (1 - eCTR) * \log(1 - y_{ctr})]$$

- Data Augmentation
  - Final stage eCTR as pseudo label for CTR task
  - Train on impression ads + non-impression ads
  - Help de-bias on both ads quality and CTR



Meta

Consolidate Ads Quality Models

- Offline Soft Recall:
  - the sum of final stage ads total value of top K ads picked by the model divided by sum of total value of the golden set.
- Total Value
  - Sum of total value for impression ads
- TVD
  - Total value divergence between final stage and early stage
- Ads quality metrics:
  - Xout rate: the ads cross-out rate
  - ASQ: a survey-assessment based metrics for ads quality related signals.

- Recall and Total Value improved
- Better Ads quality and higher CTR

Recall (+)	+3.2%
Xout rate (-)	-1.8%
ASQ (+)	+0.02
TVD (-)	-7.9%
CTR (+)	+1.7%
CVR (+)	+2.0%
Total Value (+)	+1.0%
total CPU (-)	-0.7%

Table 1: The CQS model's relative performance compared with production early stage quality models. The token (+) means better performance with higher values, and (-) means better performance with lower values.

Multi-task Learning of CQS and CTR

- Significant improvement on ads recall & total value
- CTR and CVR also increased

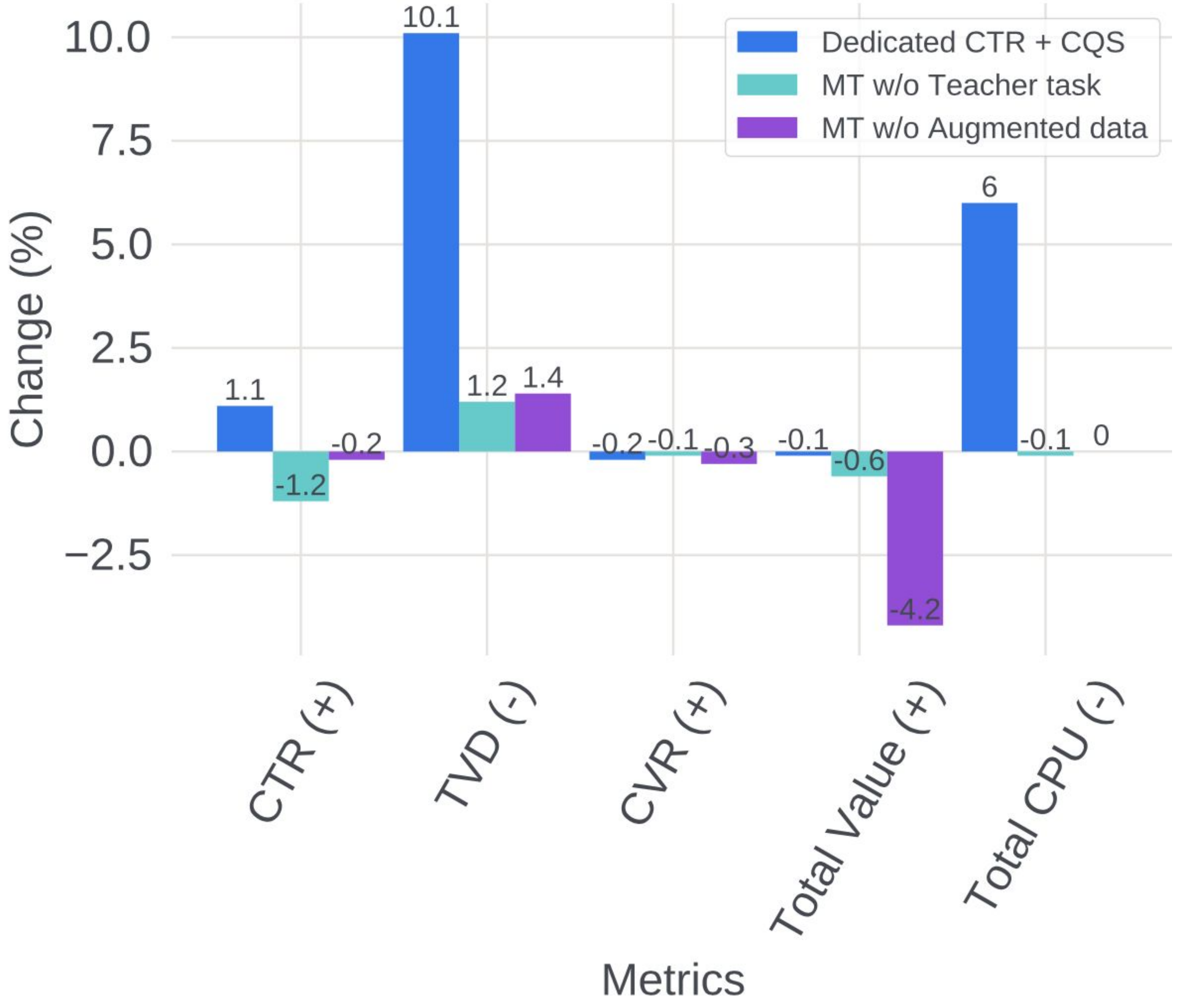
Recall (+)	+12.2%
Xout rate (-)	-3.5%
ASQ (+)	+0.005
TVD (-)	-5.7%
CTR (+)	+0.4%
CVR (+)	+0.8%
Total Value (+)	+3.0%
total CPU (-)	-0.06%

Table 2: The multi-task learning framework's relative performance compared with individual CQS model and CTR model. The token (+) means better performance with higher values, and (-) means better performance with lower values.

Ablation study

- Dedicated CTR
  - Remove CQS task in MT framework
- Dedicated CQS
  - Remove CTR & teacher task in MT framework
- MT w/o teacher:
  - Remove teacher task
- MT w/o augmented data
  - Train only on impression data

	NE diff (-)	MSE diff (-)	Recall (+)
Dedicated CTR + CQS	-0.04%	-0.6%	-0.6%
MT w/o Teacher task	+0.3%	-0.5%	-1.6%
MT w/o Augmented data	-	-	-11.9%



- Each component in our multi-task learning framework is essential to improve the performance.
- This framework can be generalized to other user cases since the CQS can be applied to any ads ranking system with the ads quality component.
- Compared with NE and MSE metrics, the offline recall evaluation metric can reflect online performance (i.e. total value) better
    - Single offline metric for an individual ranking model may not be reliable to reflect online performance.

Thank you!