Unbiased Neural Ranking Models for Product Search

About Me

- Hi I'm Laurin
- Working for Otto (otto.de)
- ML student and practitioner
- Today's talk was the topic of my Master's thesis

GMV: 6.9 billion EUR

Transition to Marketplace

4000+ vendors

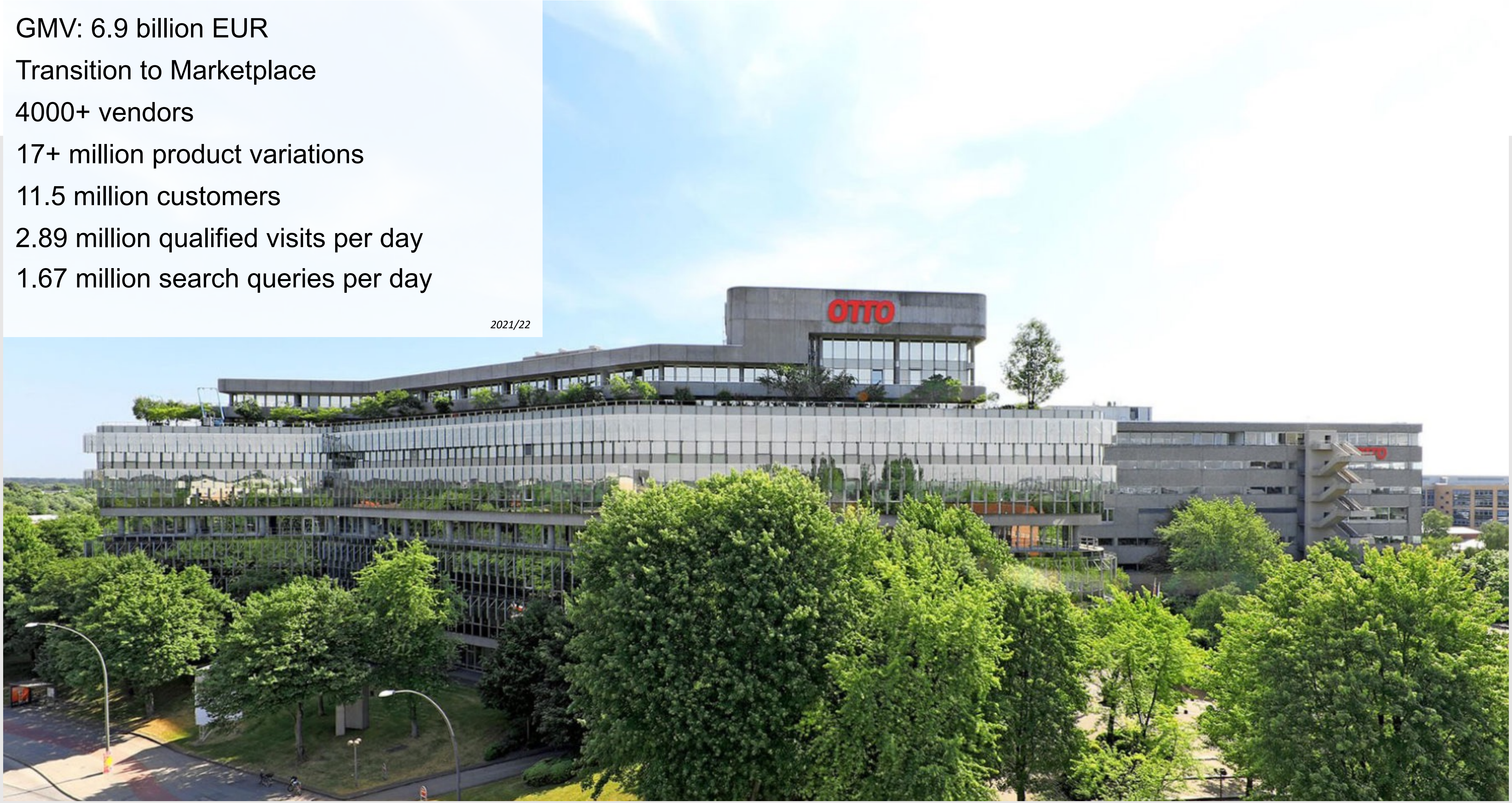
17+ million product variations

11.5 million customers

2.89 million qualified visits per day

1.67 million search queries per day

2021/22



Unbiased

(LTR)

Neural Ranking

Product Search

Unbiased

Neural Ranking

- Using Neural Networks for the ranking task
- NNs are super flexible
- Allow the use of raw text data (e.g. user query and product descriptions)

Product Search

Unbiased

- Training NNs requires tons of data
- Manually labelling data not feasible
- Alternative: Use implicit user feedback from click logs
- But: This data is biased towards current ranking (position bias)

Neural Ranking

- Using Neural Networks for the ranking task
- NNs are super flexible
- Allow the use of raw text data (e.g. user query and product descriptions)

Product Search

Unbiased

- Training NNs requires tons of data
- Manually labelling data not feasible
- Alternative: Use implicit user feedback from click logs
- But: This data is biased towards current ranking (position bias)

Neural Ranking

- Using Neural Networks for the ranking task
- NNs are super flexible
- Allow the use of raw text data (e.g. user query and product descriptions)

Product Search

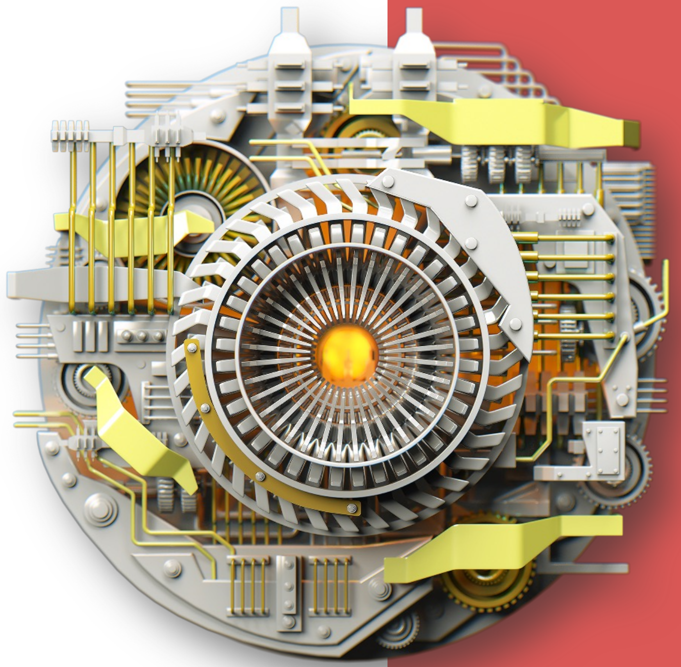
- Product search has some peculiarities:
  - Several feedback signals (clicks, purchases, a2b)
  - Several important KPIs
  - → Multi-task learning

OTTO

Neural Ranking

Why use Neural Ranking?

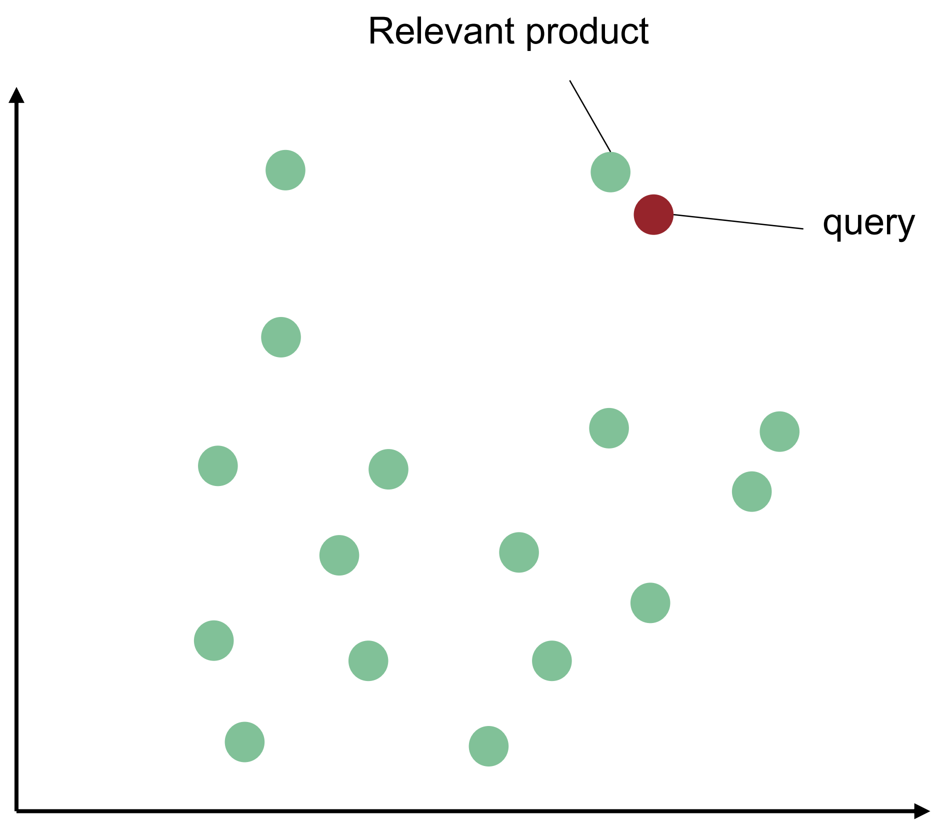
Which architectures are used?



Motivation for Neural Ranking

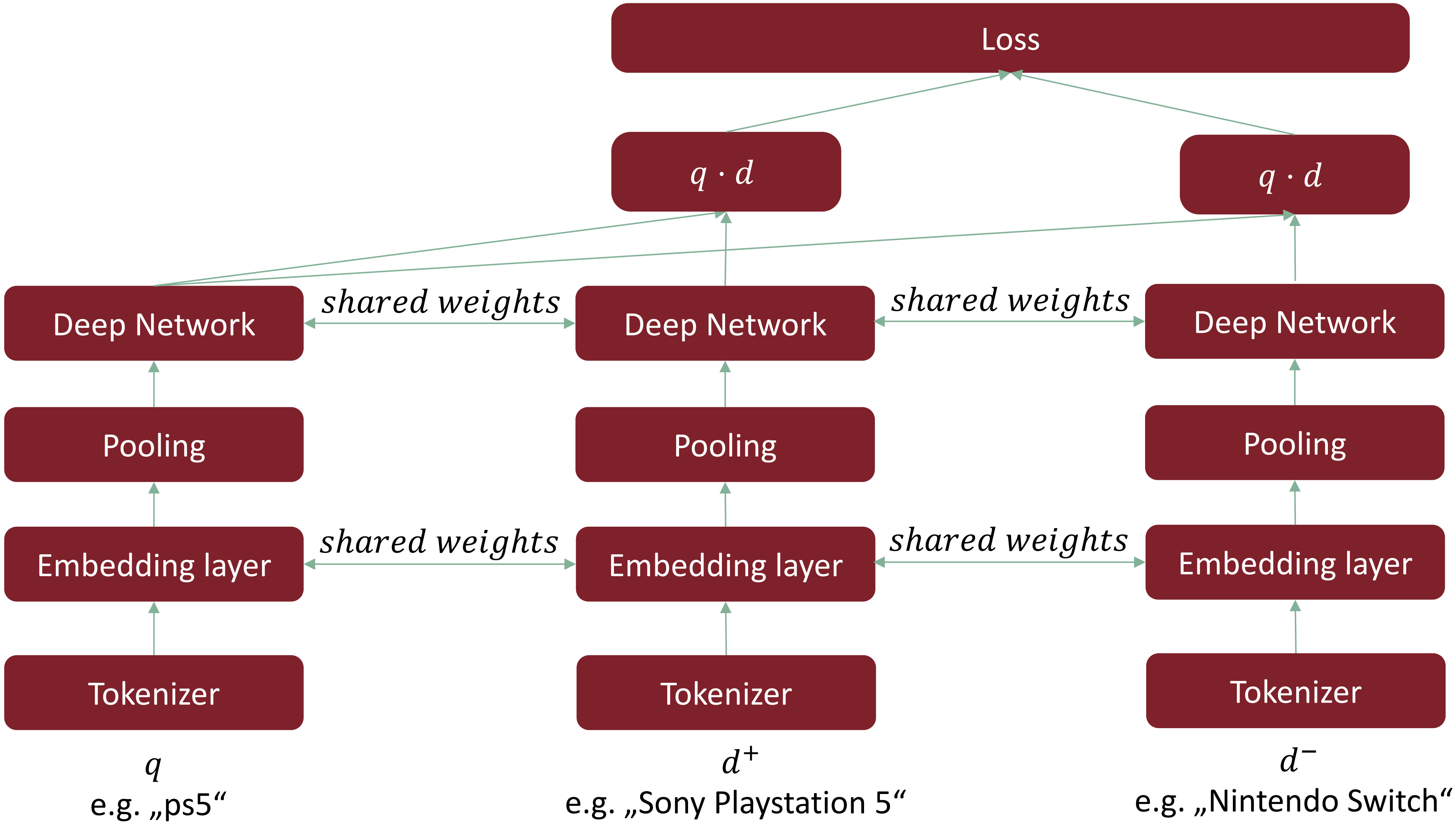
Addresses issues of exact term matching models like BM25

- Can overcome the lexical gap
    - “ps5” vs. “playstation 5”
- Learns synonyms
    - “couch” vs. “sofa”
- Can handle common spelling errors

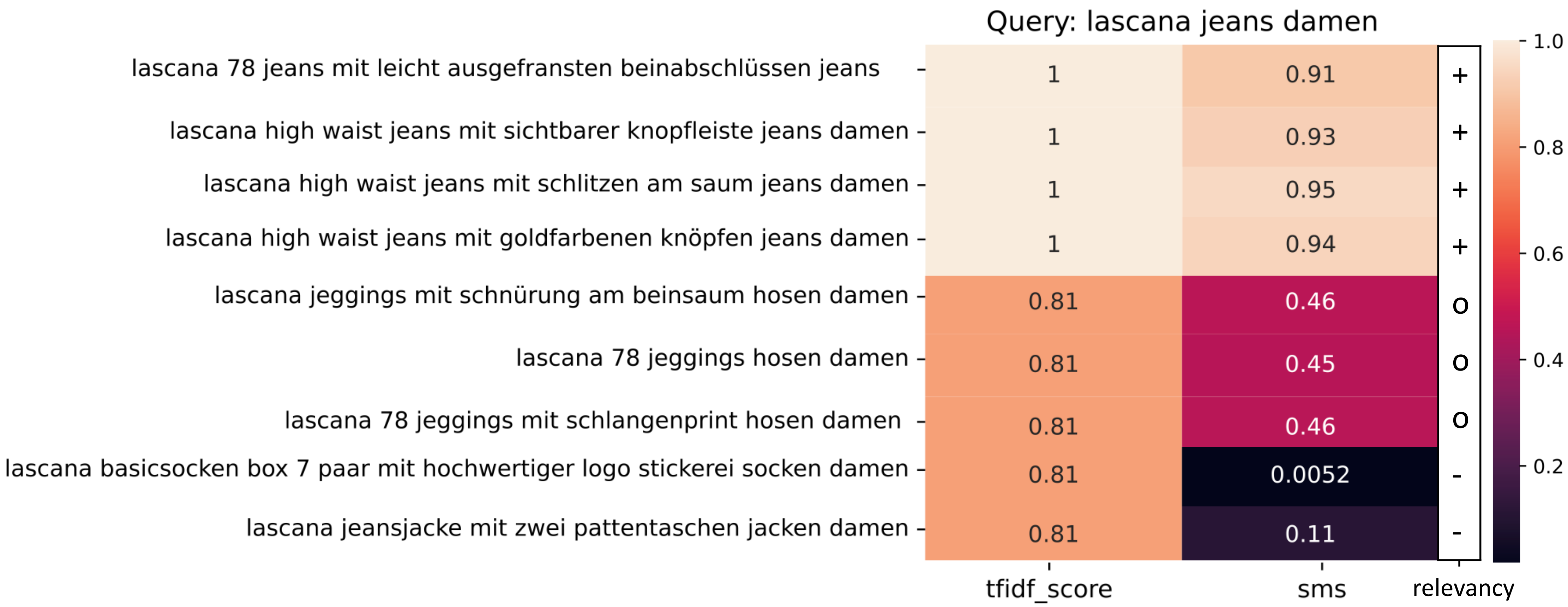


Semantic Matching

On the example of the DSSM (Huang (2013))



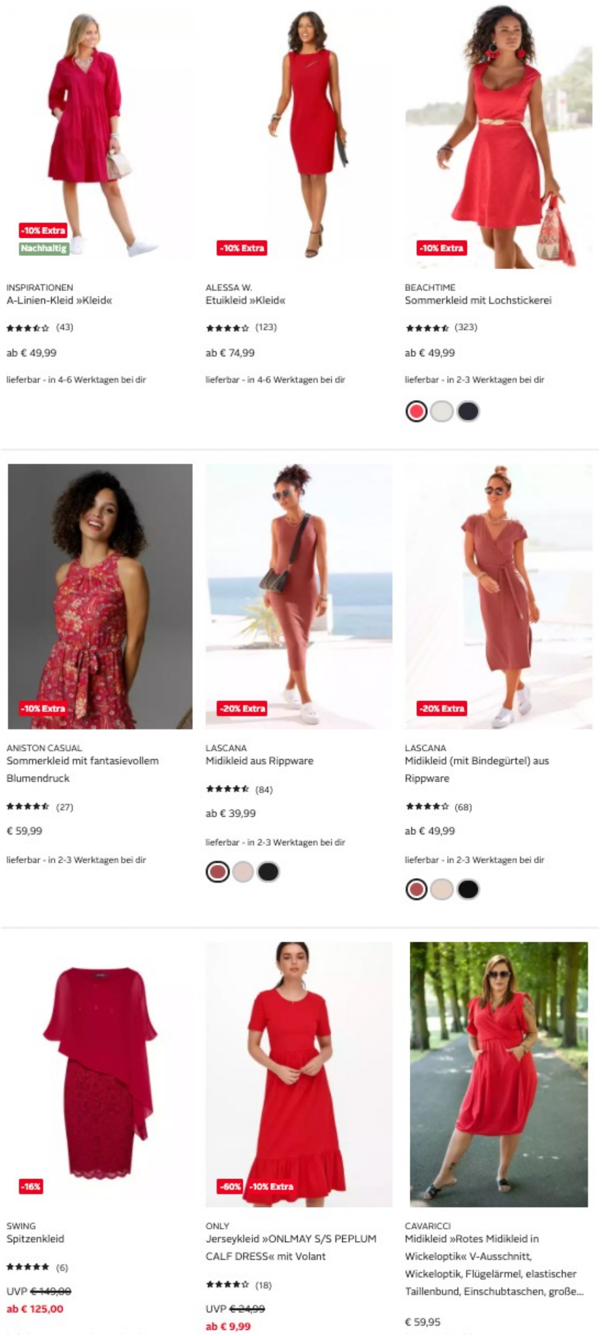
Some Intermediate Results



Semantic Matching

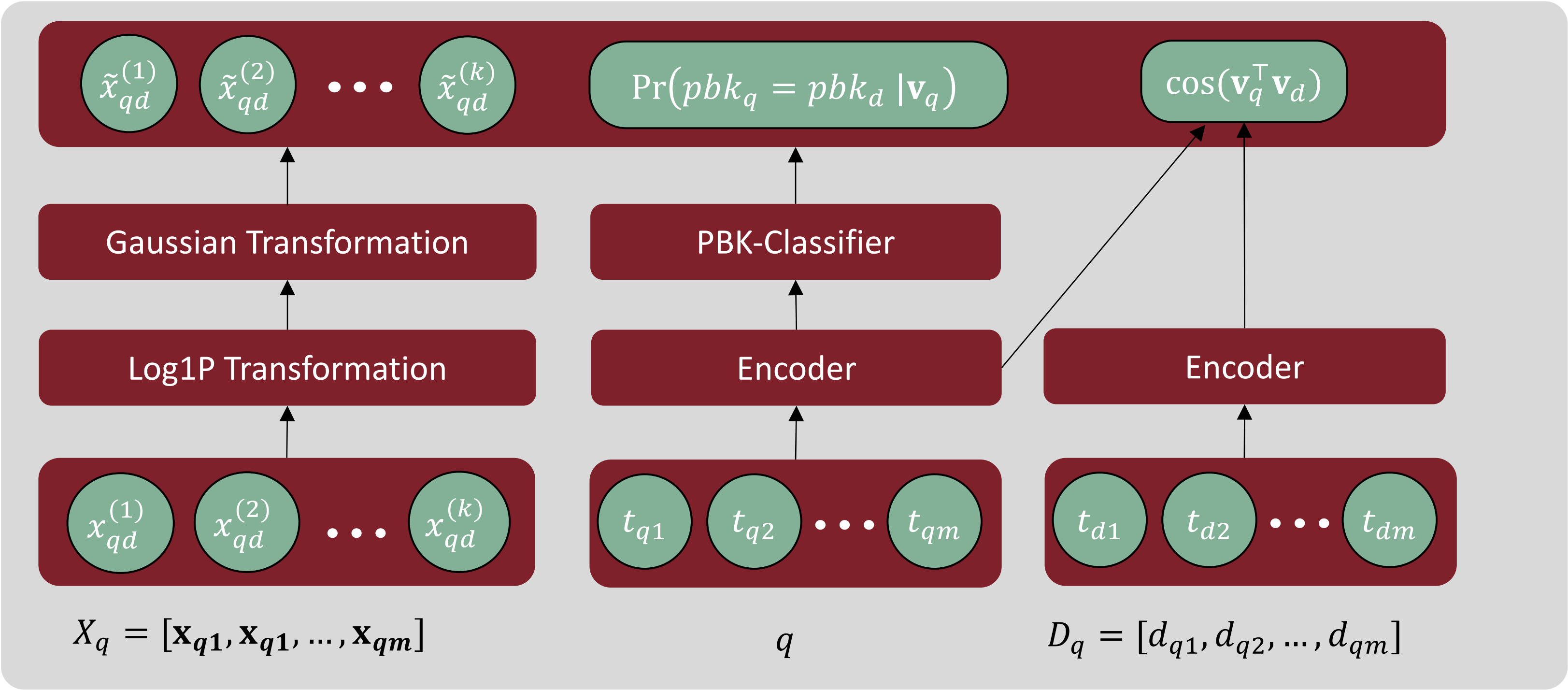
Limitations

- Semantic Matching is not equal to relevance matching
- Especially in E-Commerce there are often many perfect semantic matches to a query
- Which one is the most relevant depends on other factors
- → Semantic matching necessary but NOT sufficient condition for relevance matching

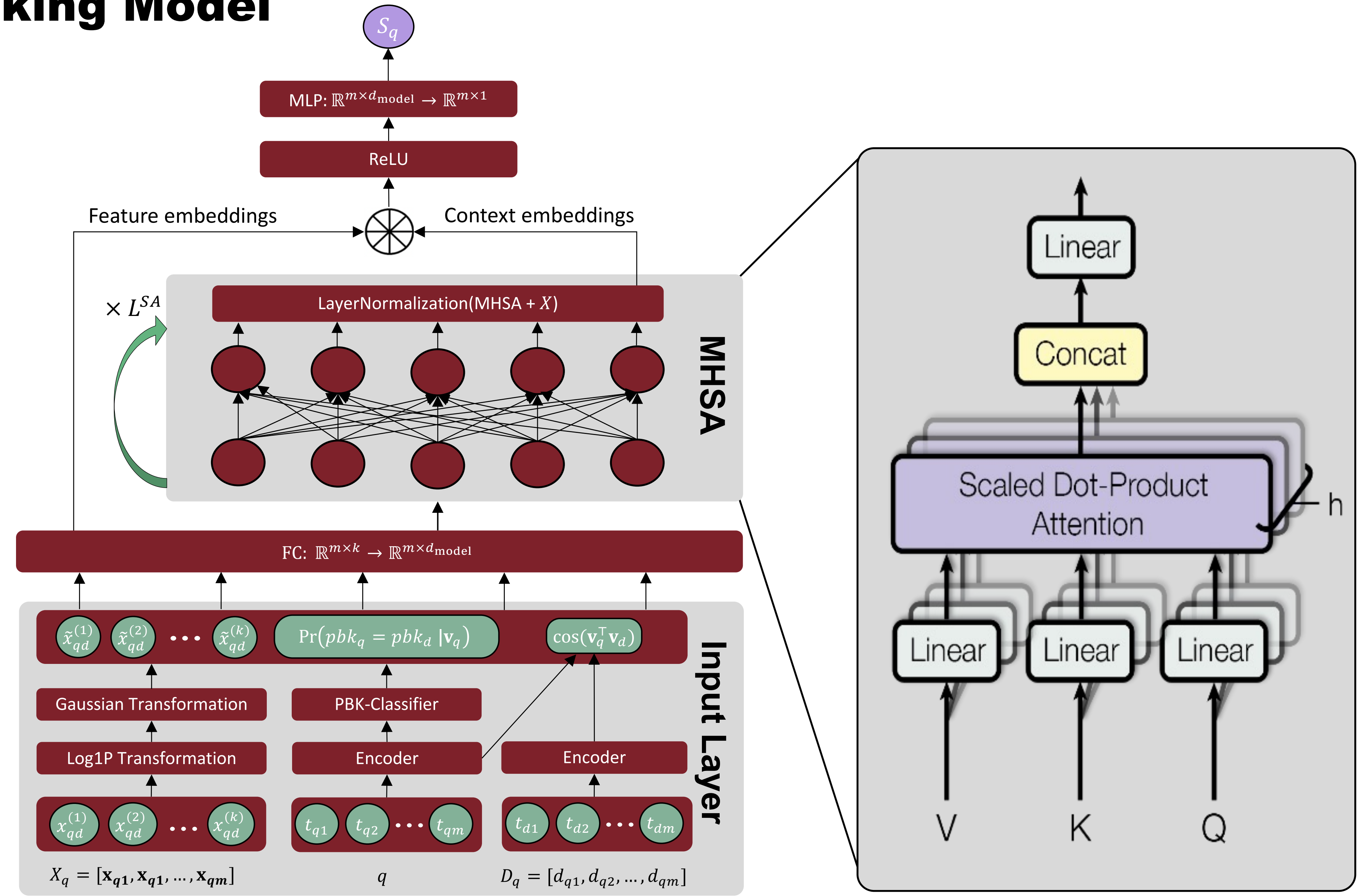


Input Space

Combining traditional LTR features with raw text matching



Ranking Model

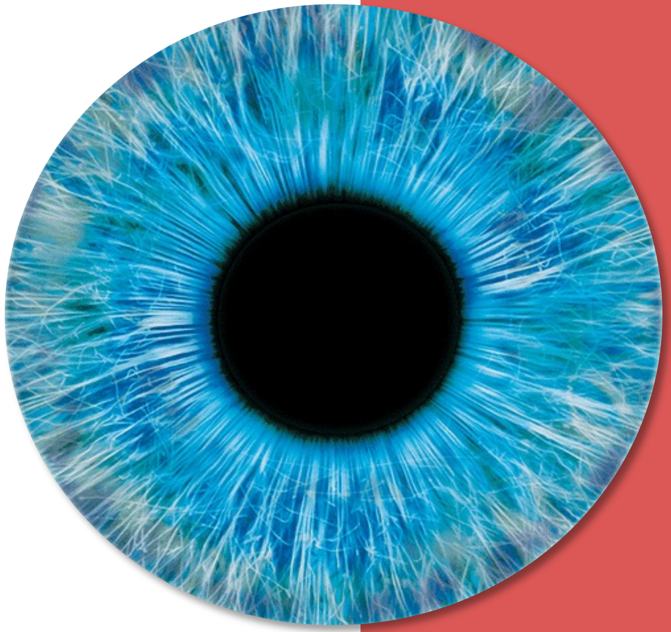


OTTO

Estimating & Isolating Position Bias

What is position bias?

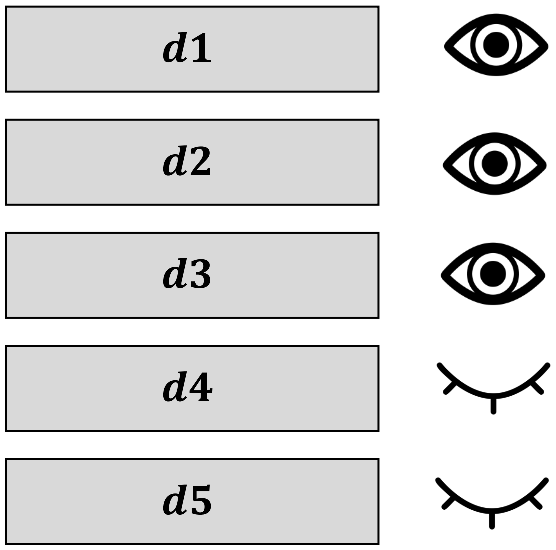
How can we deal with it?



Position Bias

The curse of low-ranked products

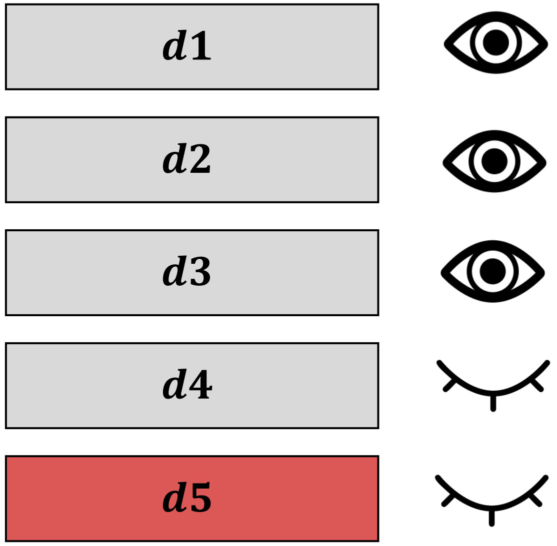
- Higher ranked products get more attention than lower-ranked ones
- Naively training on click data will not learn the product relevance



Position Bias

The curse of low-ranked products

- Higher ranked documents get more attention than lower-ranked ones
- Naively training on click data will not learn the product relevance



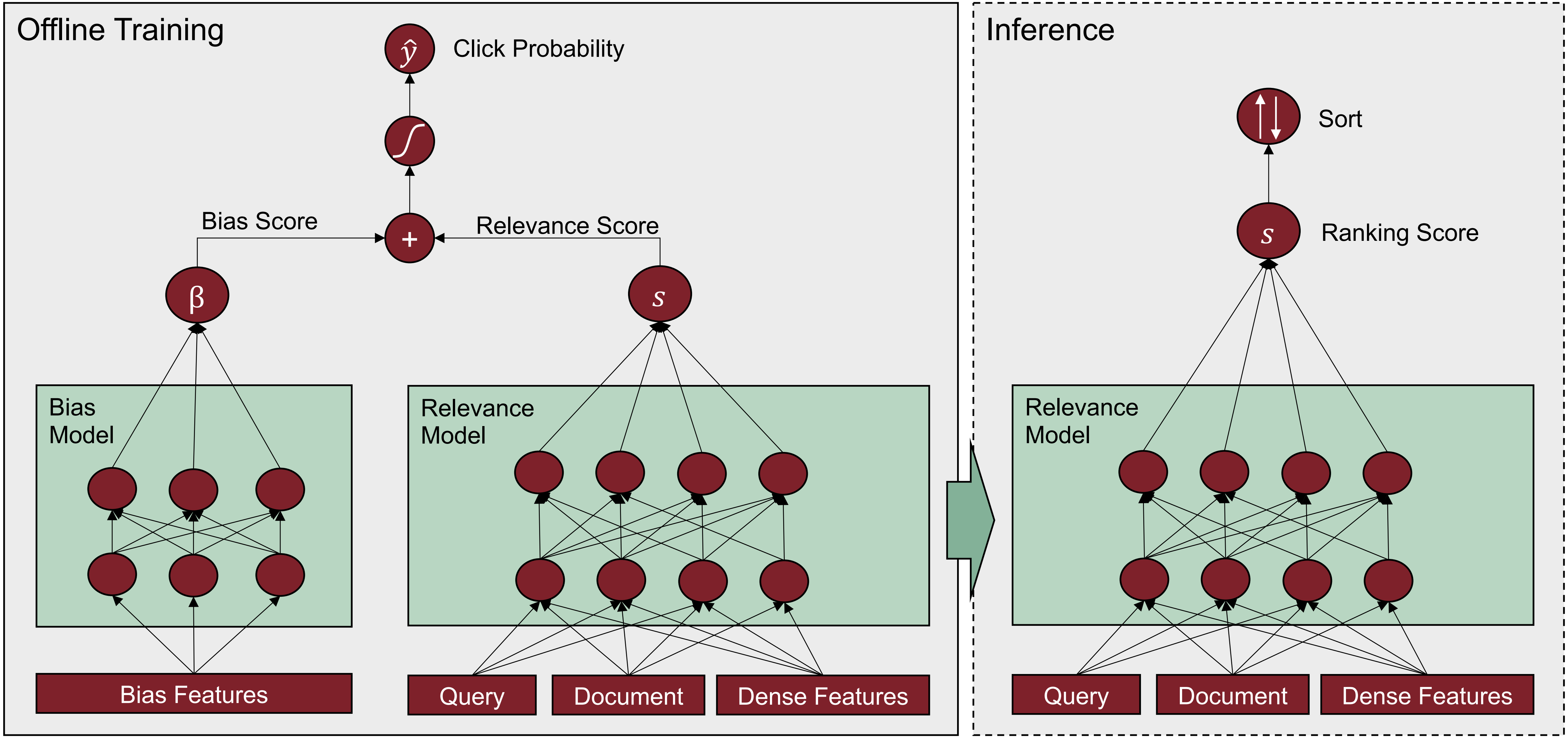
$$\Pr(C=1\,|q,d,p) = \underbrace{\Pr(E=1\,|\,p)}_{\text{examination} := \beta_p}\ \underbrace{\Pr(R=1\,|\,q,d)}_{\text{relevance} := s_{qd}}$$

Current approaches:

- Inverse Propensity Scoring
  - Requires the examination probability to be known
- EM algorithm
  - Does not yield a ranking function applicable to new (q,d)-pairs

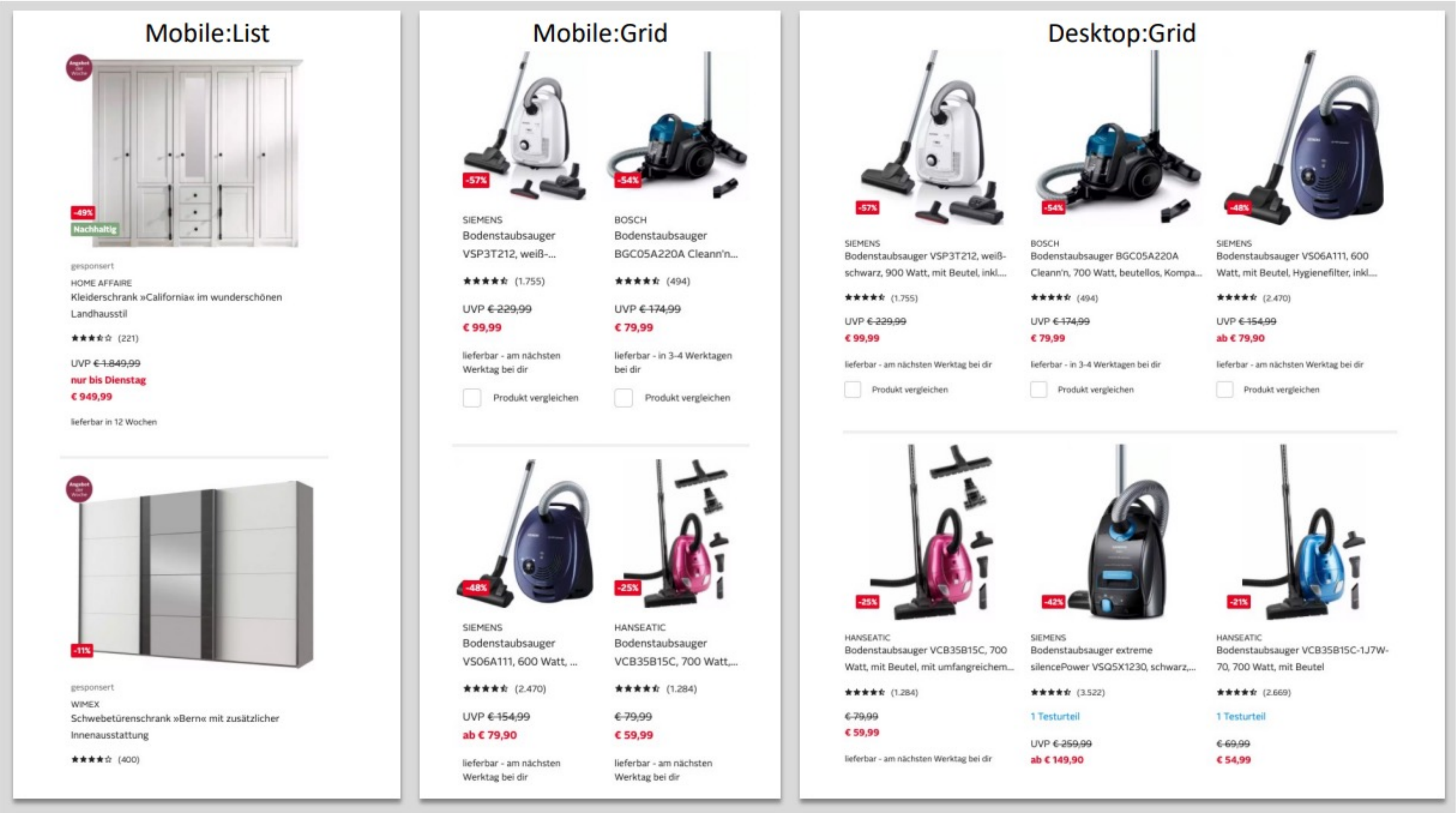
Jointly Estimating Relevance and Position Bias

Our „JoE“ approach



Different Layouts in E-Commerce

Mean also different position biases?

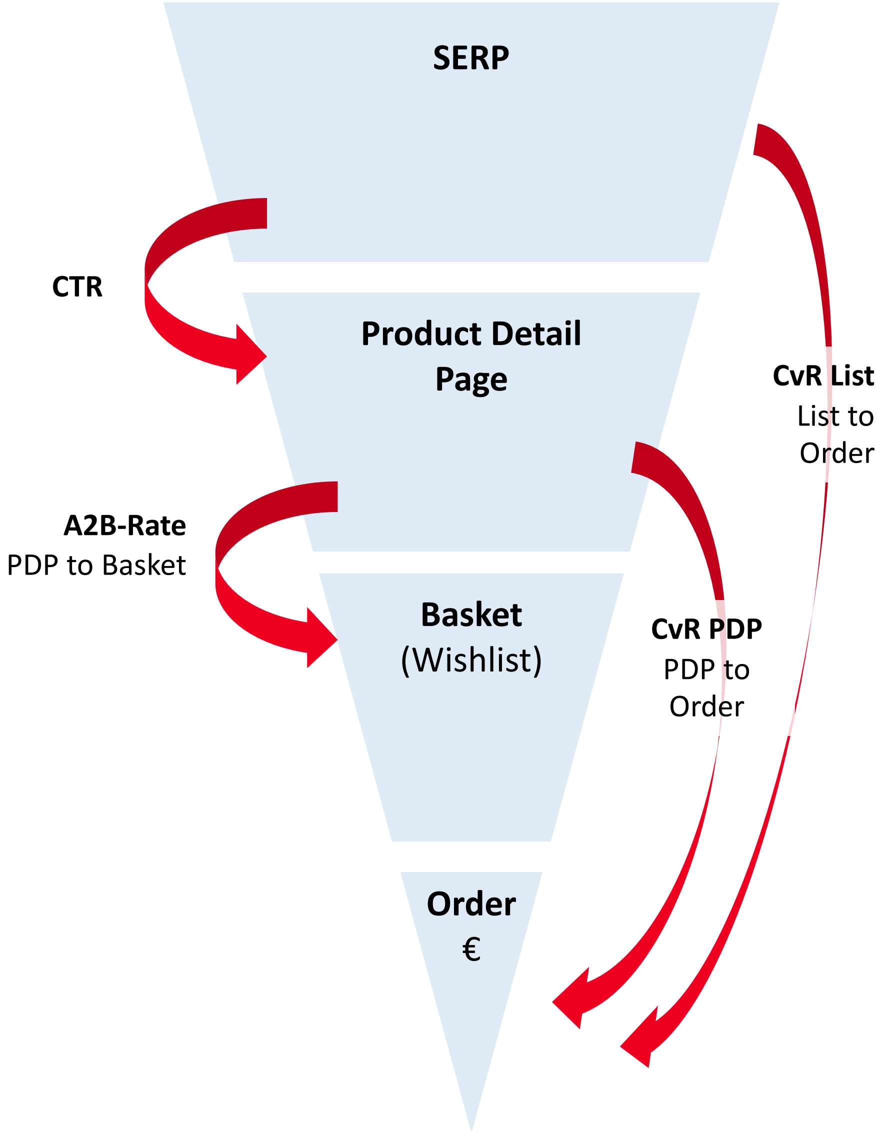


Multi-Task Learning

In Product Search

What is special about product search?

How can we account for it?

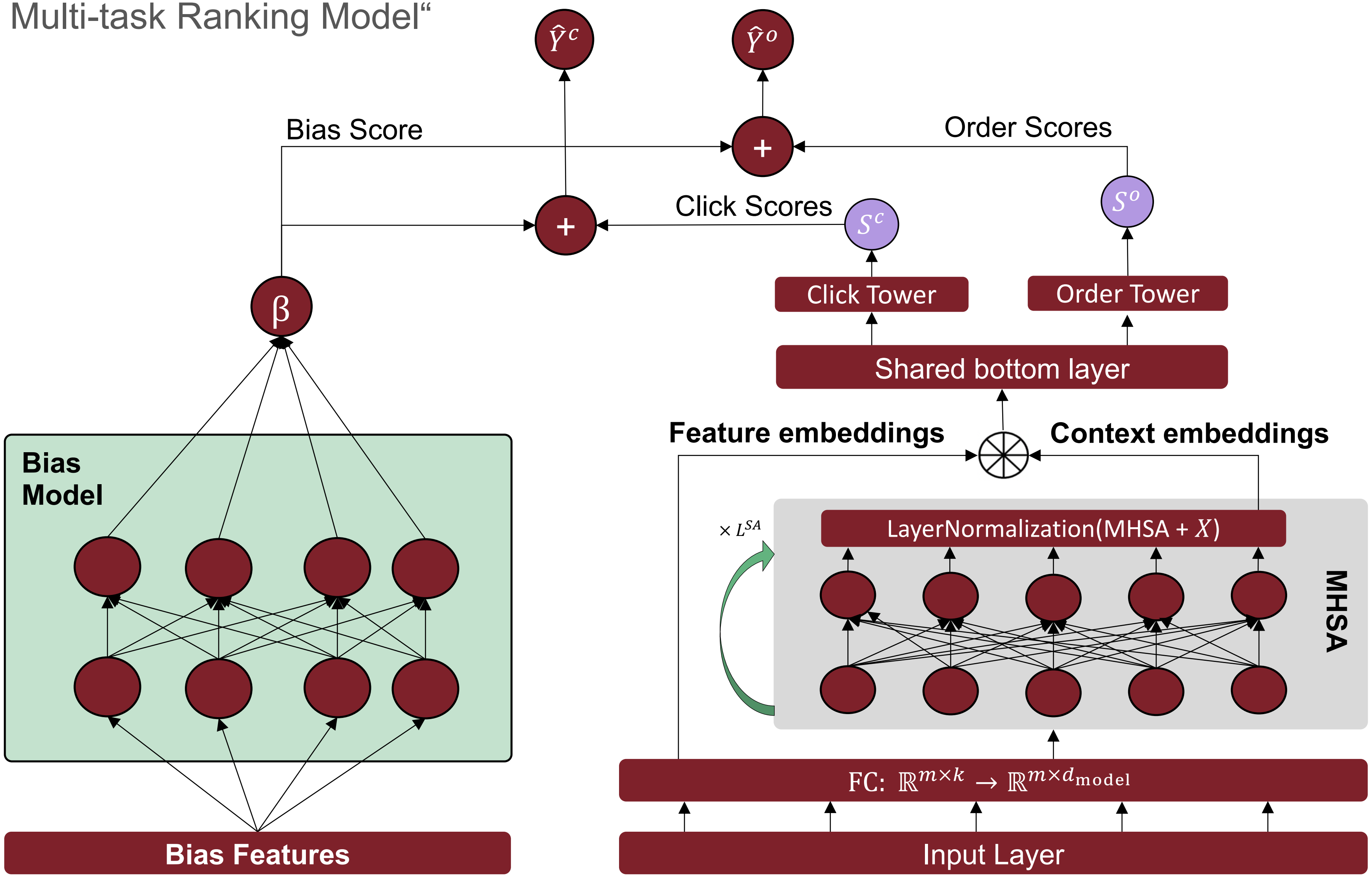


- Customers have different intents (e.g. browsing and shopping)
- We want to serve these intents and optimise different KPIs
  - CTR, CvR, GMV
- We can track the whole funnel
  - Clicks, Add2Baskets, Purchases

OTTO

The final Model

„Unbiased Multi-task Ranking Model“

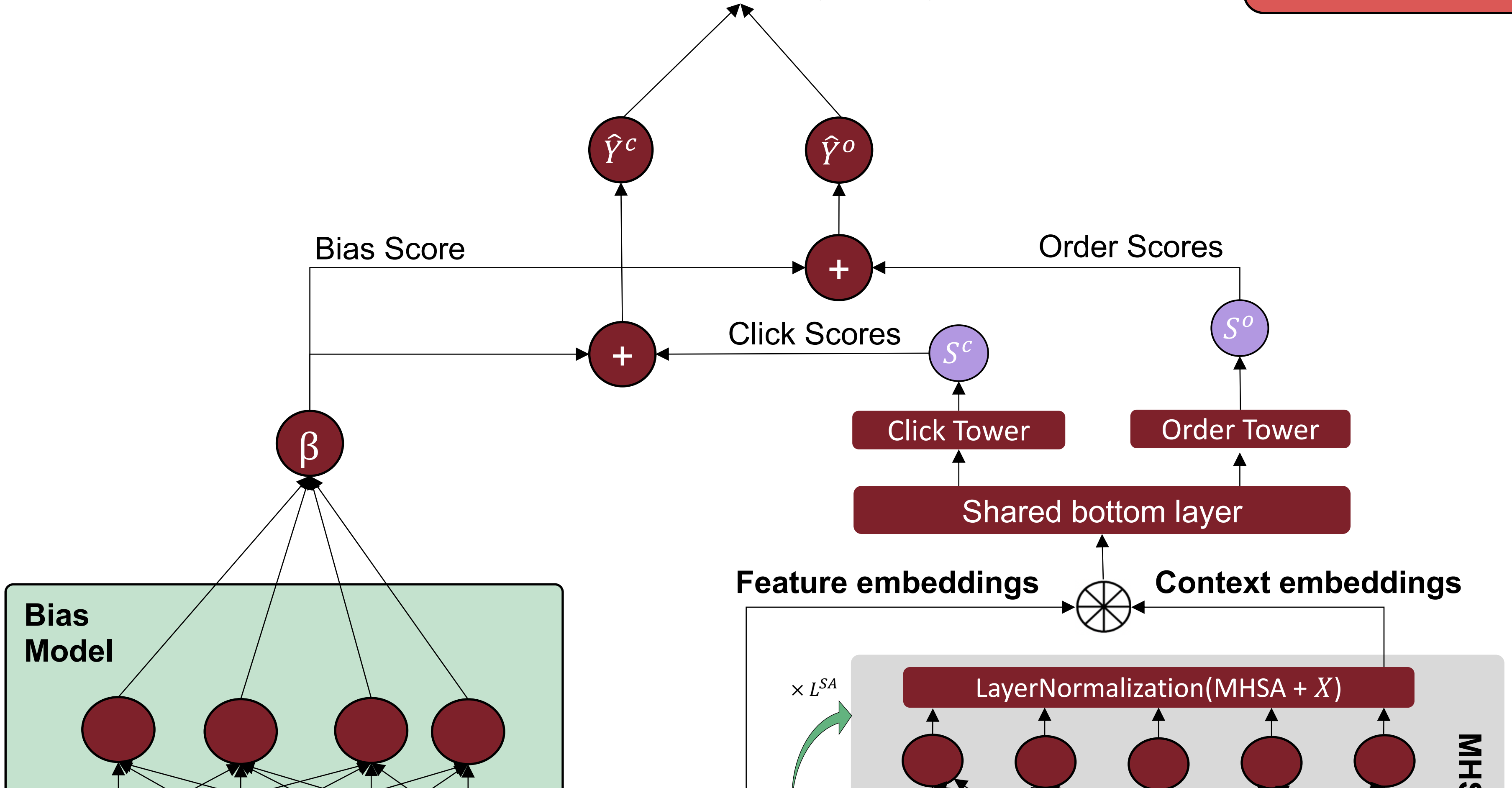


The final Model

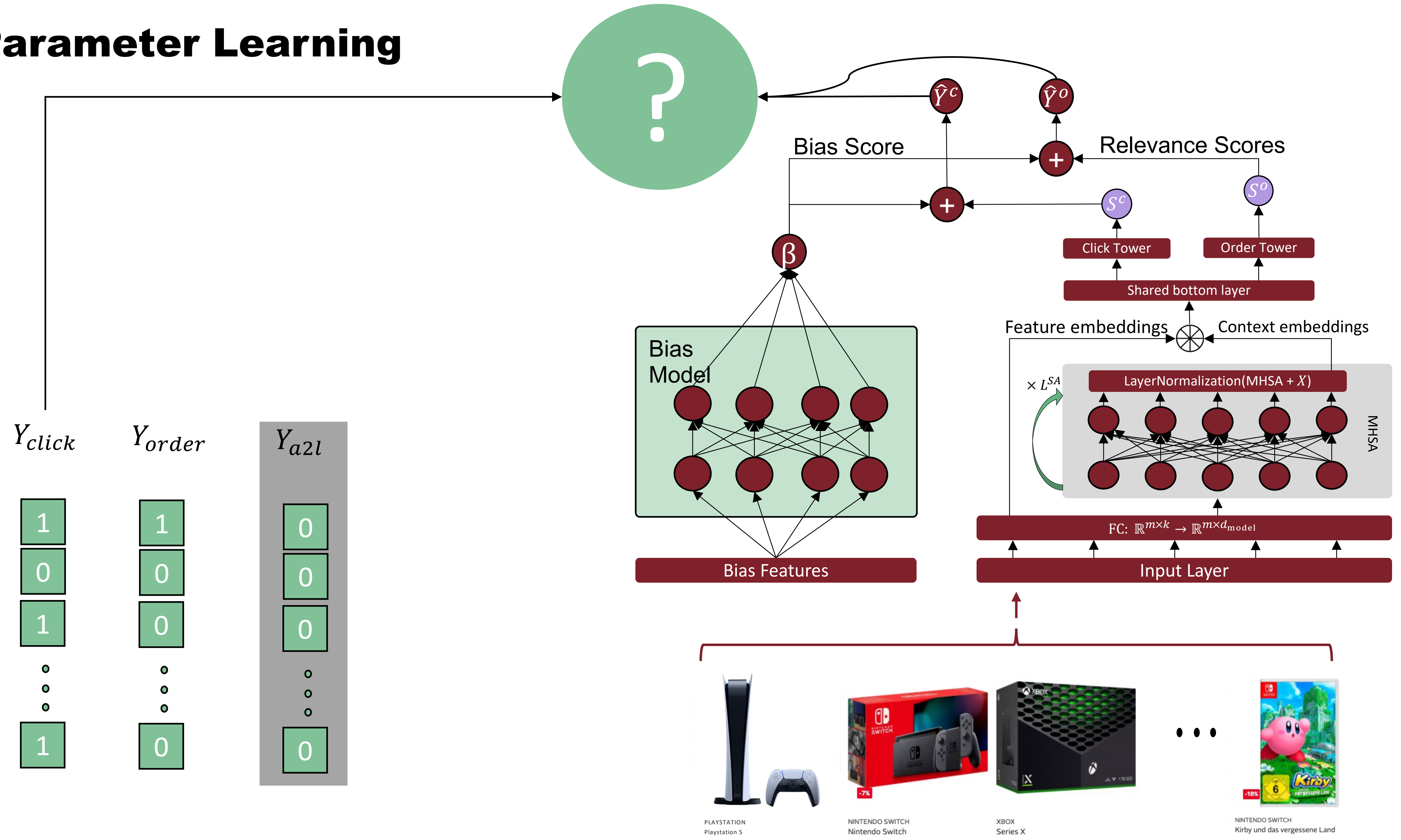
„Unbiased Multi-task Ranking Model“

$$\hat{Y}^{final} = \alpha \cdot \hat{Y}^{click} + (1-\alpha)\hat{Y}^{order}$$

Choose α so as to optimize your online metrics



Parameter Learning



Learning contd.

- This is in particular a multi-label classification problem (each example (SERP) can have clicks on multiple items or none at all)
- In multilabel classification, we typically use sigmoid cross entropy (CE) loss

$$f(s_i) = \frac{1}{1 + e^{-s_i}} \qquad \ell = -\sum_i y_i \log(f(s_i))$$

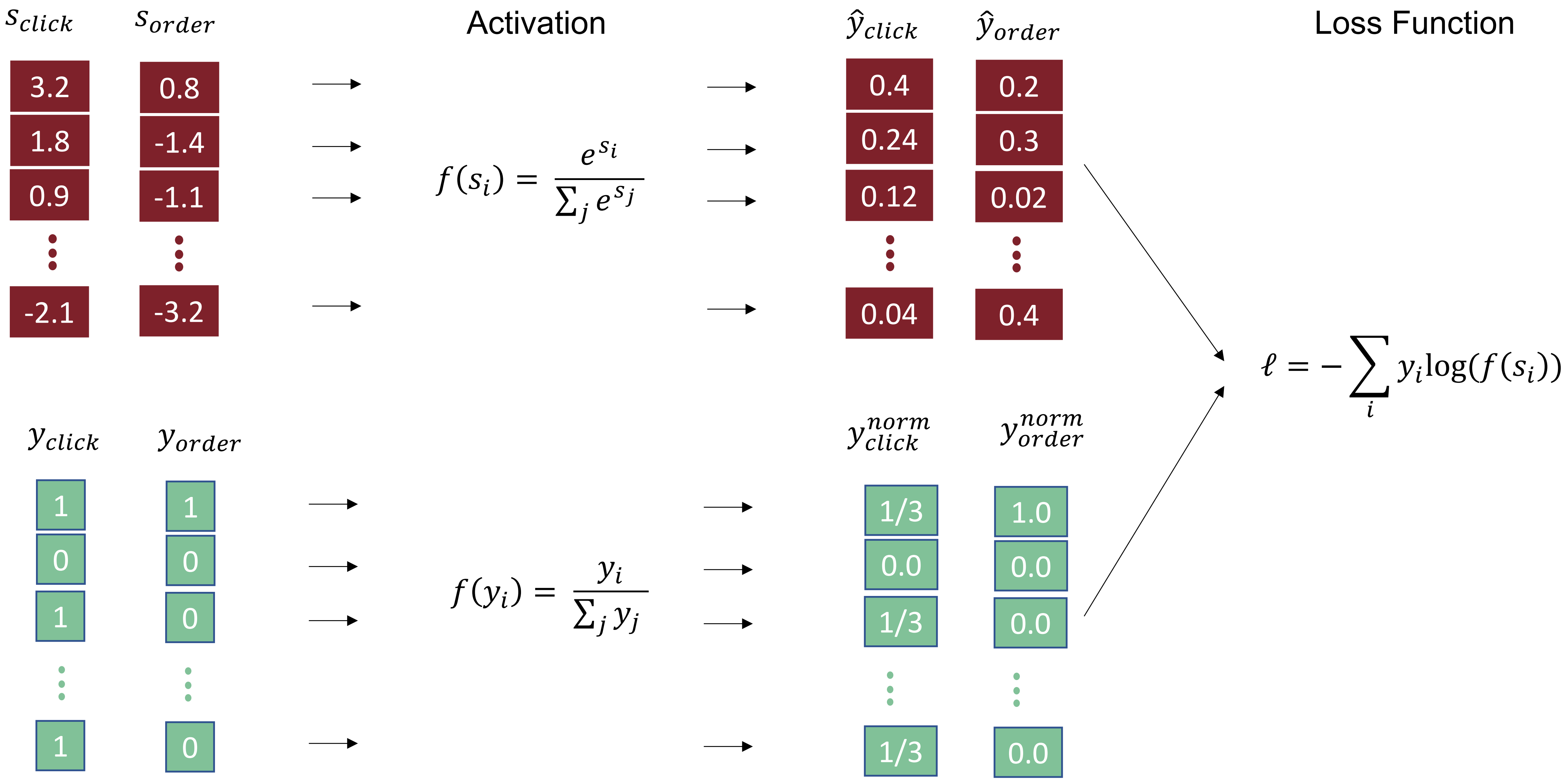
- BUT: sigmoid CE does not necessarily approximate ranking metrics well

- Bruch et al. (from Google research) show „that **softmax cross entropy is a bound on Mean Reciprocal Rank (MRR) as well as NDCG** when working with binary ground-truth labels“

$$f(s_i) = \frac{e^{s_i}}{\sum_j e^{s_j}} \qquad \ell = -\sum_i y_i \log(f(s_i))$$

- Another problem: sum over softmax equals to one, but sum over labels unlikely to sum to one!
- Normalize labels by their sum

Learning contd.

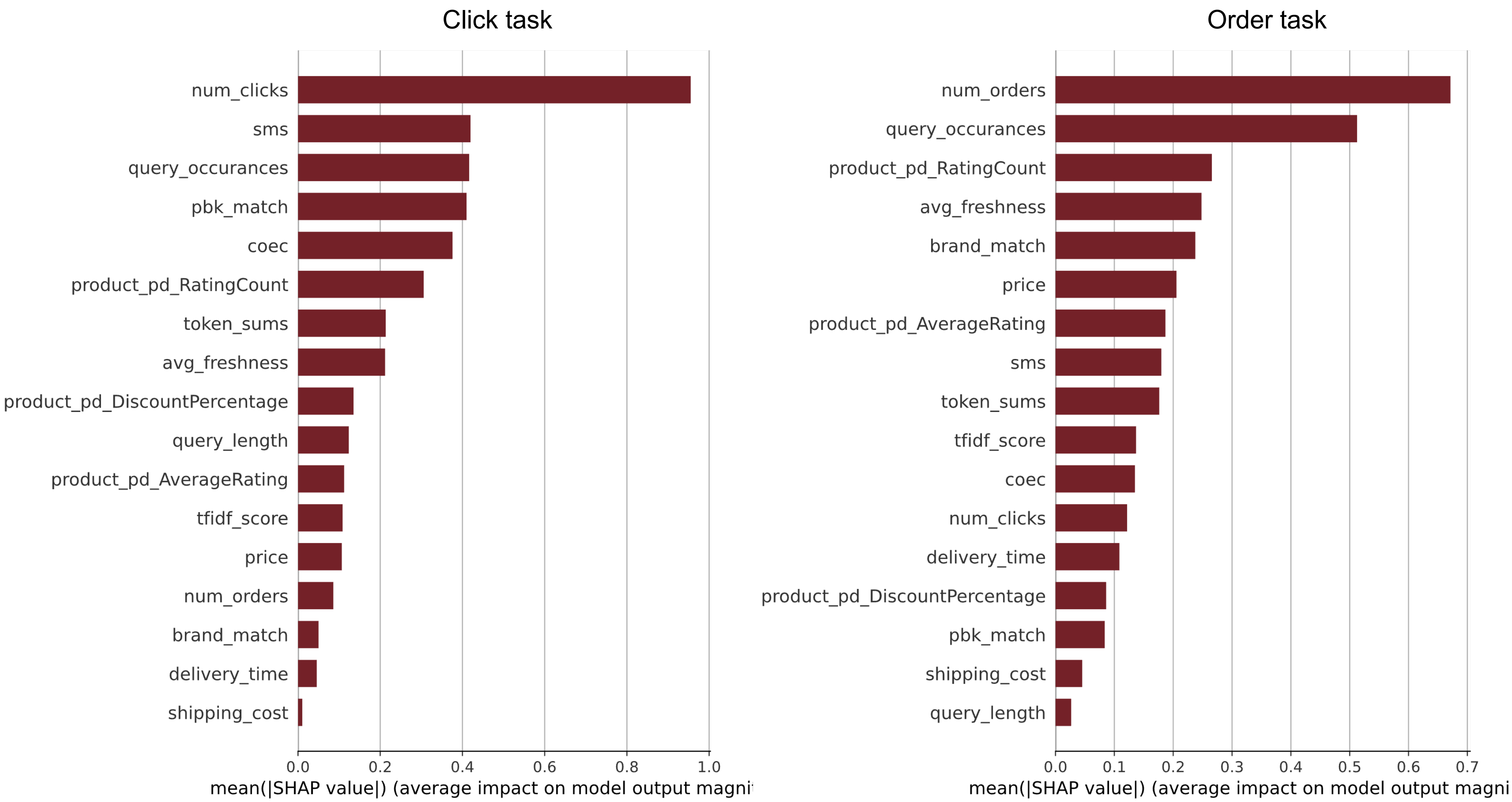


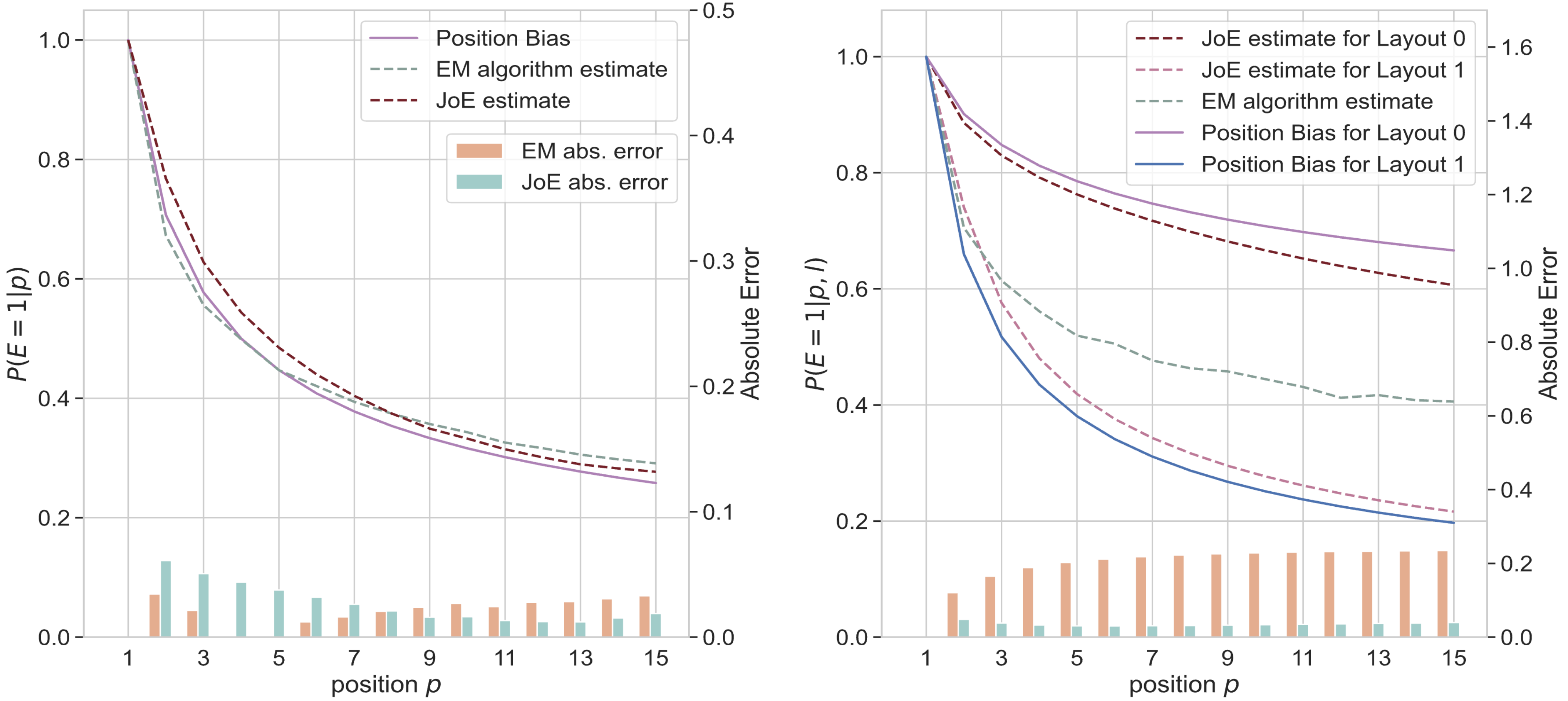
Results

What did we observe offline?

Were we able to improve online?

Feature Importances





Position Bias
EM algorithm estimate
JoE estimate
EM abs. error
JoE abs. error
$P(E = 1|p)$
Absolute Error
position p
JoE estimate for Layout 0
JoE estimate for Layout 1
EM algorithm estimate
Position Bias for Layout 0
Position Bias for Layout 1
$P(E = 1|p, l)$
Absolute Error
position p

CTR

▲0.7%

ranking CTR List --> qADS (Session) : E651A

CTR@10

▲3.3%

ranking CTR List --> qADS pos 1-10 organic (Session) : E651A

CTR@3

▲5.6%

ranking CTR List --> qADS pos 1-3 organic (Session) : E651A

PDP to Order

▼0.1%

PDP to A2B

▼3.1%

A2B to Units

▲0.3%

Conversion Rate

▲0.7%

CvR List (Session) : E651A

Orders

▲0.9%

Orders OTTO

▲0.5%

Orders PARTNER

▲1.7%

Orders MIXED

▲1.8%

Units

▲0.2%

units : E651A

Revenue

▲0.3%

revenue [orders less 5.000 EUR] : E651A

OTTO

Thanks for your attention

Any questions?

…and by the way: **we are hiring**!

(https://www.otto.de/jobs/jobsuche/search/)