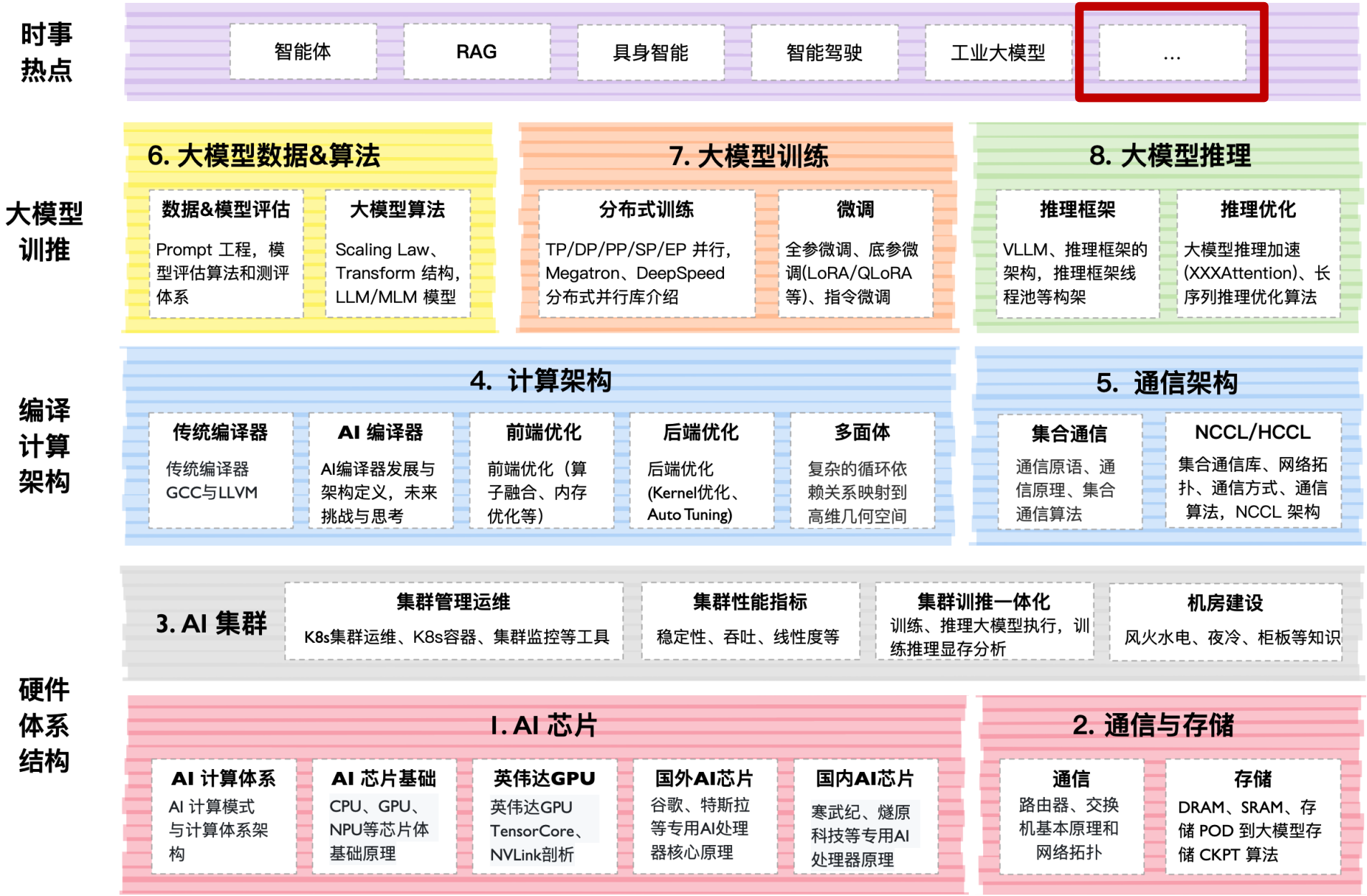
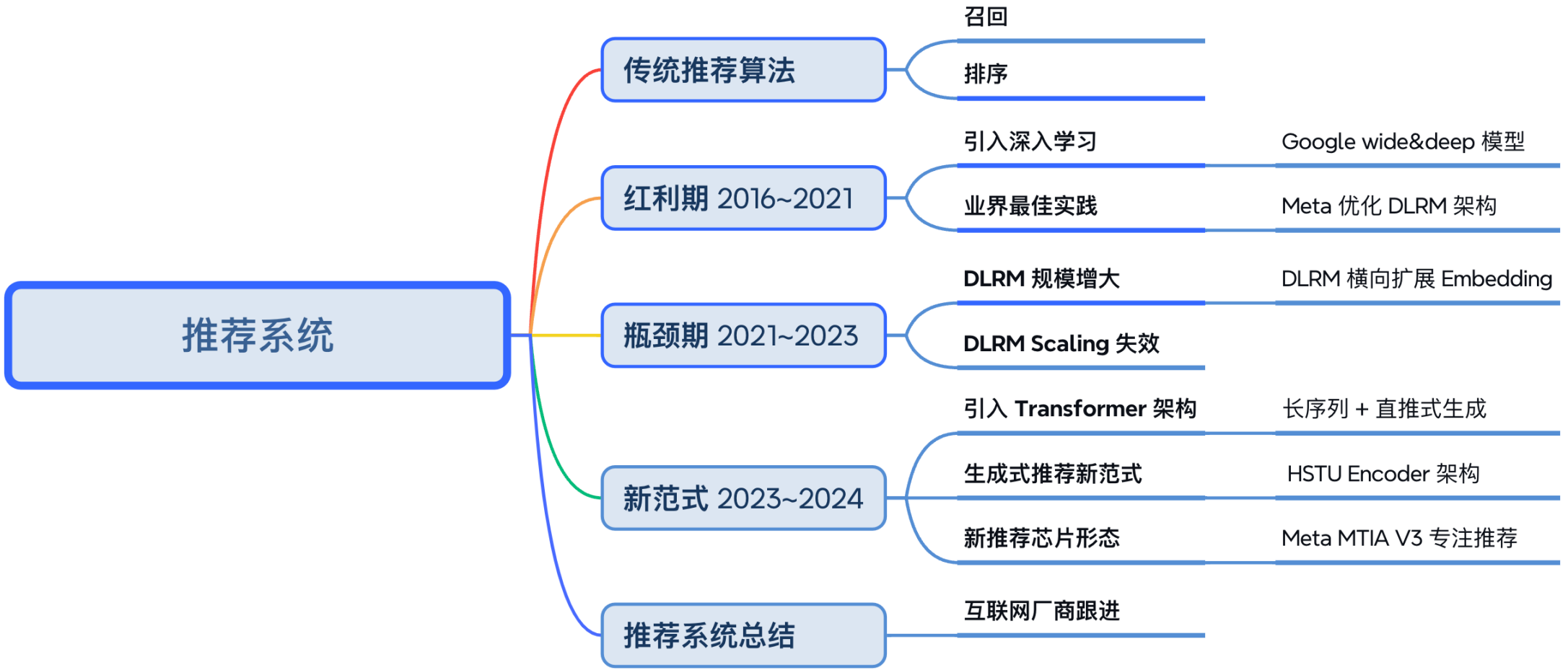

BATCH PROCESSING
REAL-TIME ADAPTATION
POPULARITY-DRIVEN SUGGESTIONS
CONTEXT-DRIVEN PERSONALIZATION
REGULAR UPDATES NEEDE
生成式推荐
算法解读
GEN AI RECOMMENDATION ENGINES
FOR ECOMMERCE PERSONALIZATION

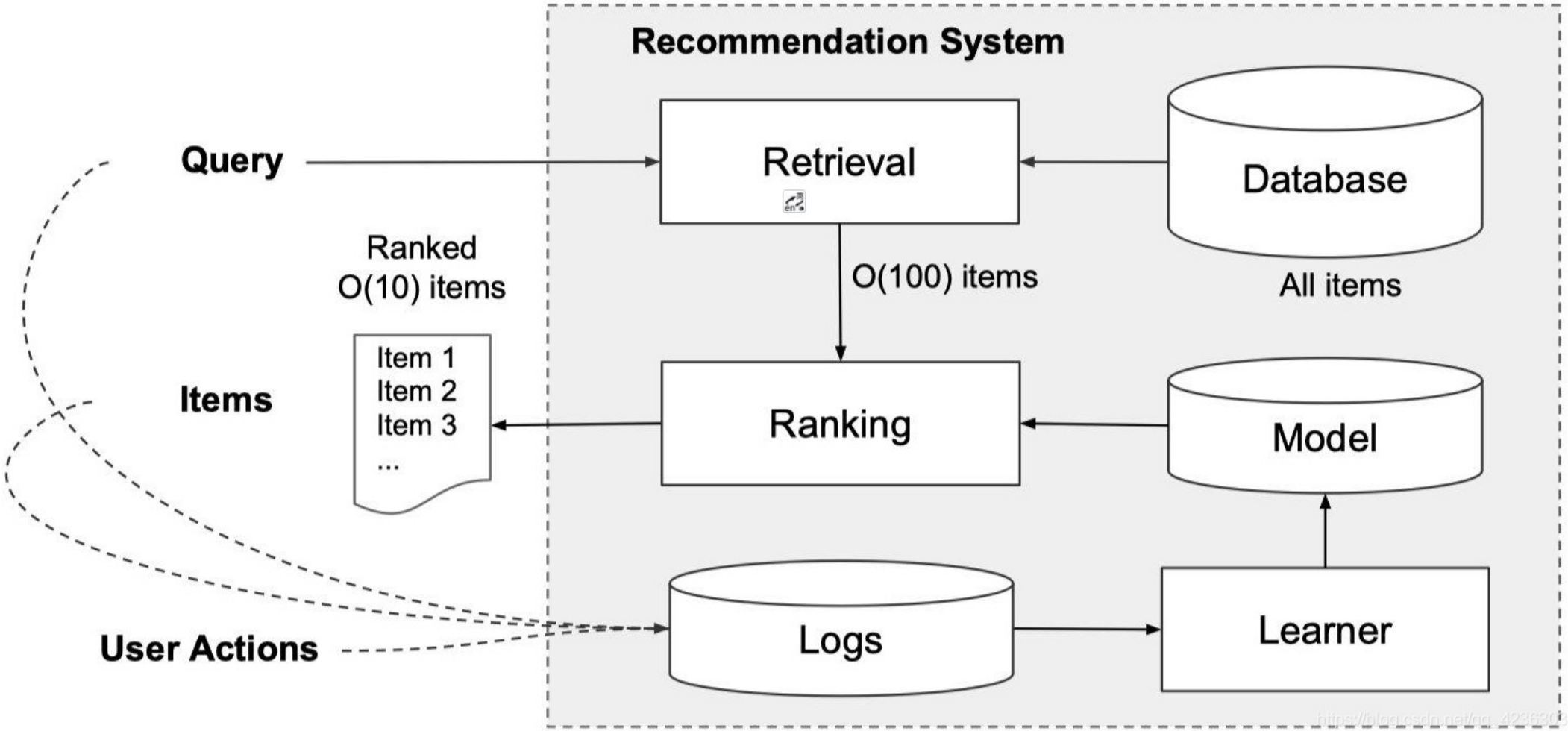

时事热点
智能体
RAG
具身智能
智能驾驶
工业大模型
...
大模型训推
6. 大模型数据&算法
数据&模型评估
Prompt 工程，模型评估算法和测评体系
大模型算法
Scaling Law、Transform 结构，LLM/MLM 模型
7. 大模型训练
分布式训练
TP/DP/PP/SP/EP 并行，Megatron、DeepSpeed 分布式并行库介绍
微调
全参微调、底参微调(LoRA/QLoRA 等)、指令微调
8. 大模型推理
推理框架
VLLM、推理框架的架构，推理框架线程池等构架
推理优化
大模型推理加速(XXXAttention)、长序列推理优化算法
编译计算架构
4. 计算架构
传统编译器
传统编译器 GCC与LLVM
AI 编译器
AI编译器发展与架构定义，未来挑战与思考
前端优化
前端优化（算子融合、内存优化等）
后端优化
后端优化(Kernel优化、Auto Tuning)
多面体
复杂的循环依赖关系映射到高维几何空间
5. 通信架构
集合通信
通信原语、通信原理、集合通信算法
NCCL/HCCL
集合通信库、网络拓扑、通信方式、通信算法，NCCL 架构
硬件体系结构
3. AI 集群
集群管理运维
K8s集群运维、K8s容器、集群监控等工具
集群性能指标
稳定性、吞吐、线性度等
集群训推一体化
训练、推理大模型执行，训练推理显存分析
机房建设
风火水电、夜冷、柜板等知识
1. AI 芯片
AI 计算体系
AI 计算模式与计算体系架构
AI 芯片基础
CPU、GPU、NPU等芯片体基础原理
英伟达GPU
英伟达GPU TensorCore、NVLink剖析
国外AI芯片
谷歌、特斯拉等专用AI处理器核心原理
国内AI芯片
寒武纪、燧原科技等专用AI处理器原理
2. 通信与存储
通信
路由器、交换机基本原理和网络拓扑
存储
DRAM、SRAM、存储 POD 到大模型存储 CKPT 算法

什么是生成式推荐


推荐系统
传统推荐算法
召回
排序
红利期 2016~2021
引入深入学习
Google wide&deep 模型
业界最佳实践
Meta 优化 DLRM 架构
瓶颈期 2021~2023
DLRM 规模增大
DLRM 横向扩展 Embedding
DLRM Scaling 失效
新范式 2023~2024
引入 Transformer 架构
长序列 + 直推式生成
生成式推荐新范式
HSTU Encoder 架构
新推荐芯片形态
Meta MTIA V3 专注推荐
推荐系统总结
互联网厂商跟进

01 传统推荐算法

推荐系统的架构图


Recommendation System
Query
Retrieval
Database
Ranked
O(10) items
O(100) items
All items
Items
Item 1
Item 2
Item 3
...
Ranking
Model
User Actions
Logs
Learner

推荐系统2个阶段：召回 + 排序

阶段	特点	特点
召回	根据用户部分特征，从海量的物品库里，快速找回一小部分用户潜在感兴趣的物品	速度快
排序	可以融入较多特征，使用复杂模型，来精准地做个性化推荐	结果精准

推荐系统4个阶段：精细划分

阶段	特点
召回	从海量物品中快速找回一部分重要物品
粗排	进行粗略排序，保证一定精准度并减少物品数量
精排	精准地对物品进行个性化排序
重排	改进用户体验

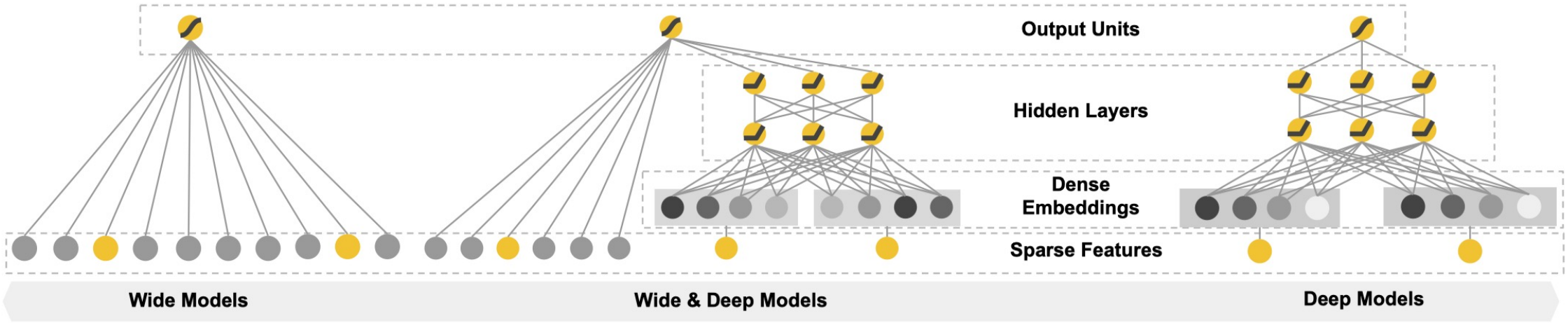
02 红利期

2016~2021

Deep learning recommendation model for personalization and recommendation systems

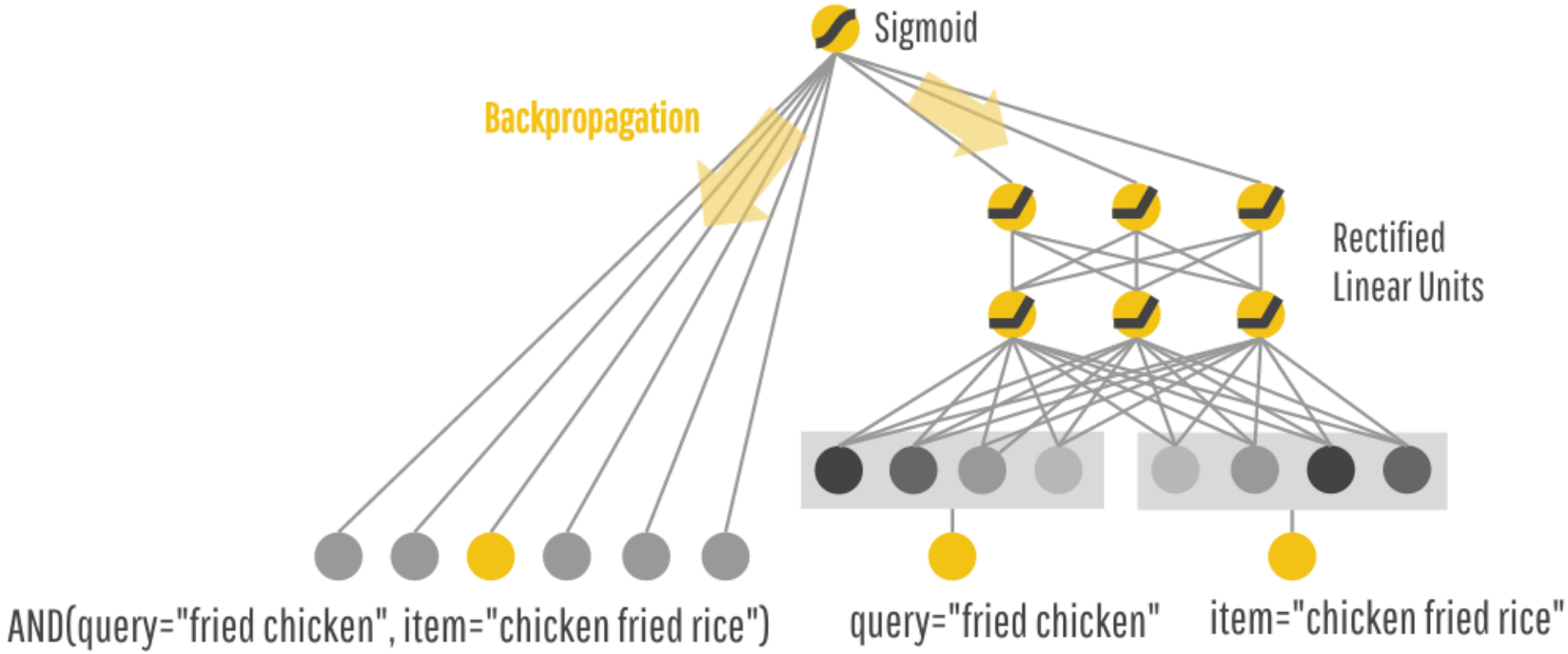
1. wide&deep 模型

- wide&deep 模型讨论如何利用深度学习模型来进行推荐系统的CTR预测，是在推荐系统一次深度学习的成功尝试。
- 让模型兼具逻辑回归和深度神经网络的优点，既能快速处理和记忆大量历史行为特征，又具有强大的表达能力；



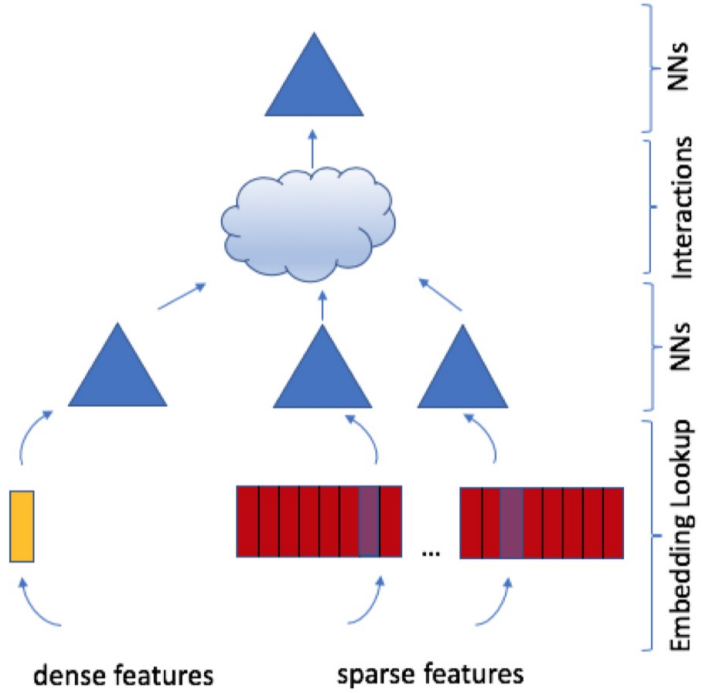
2. wide&deep 模型

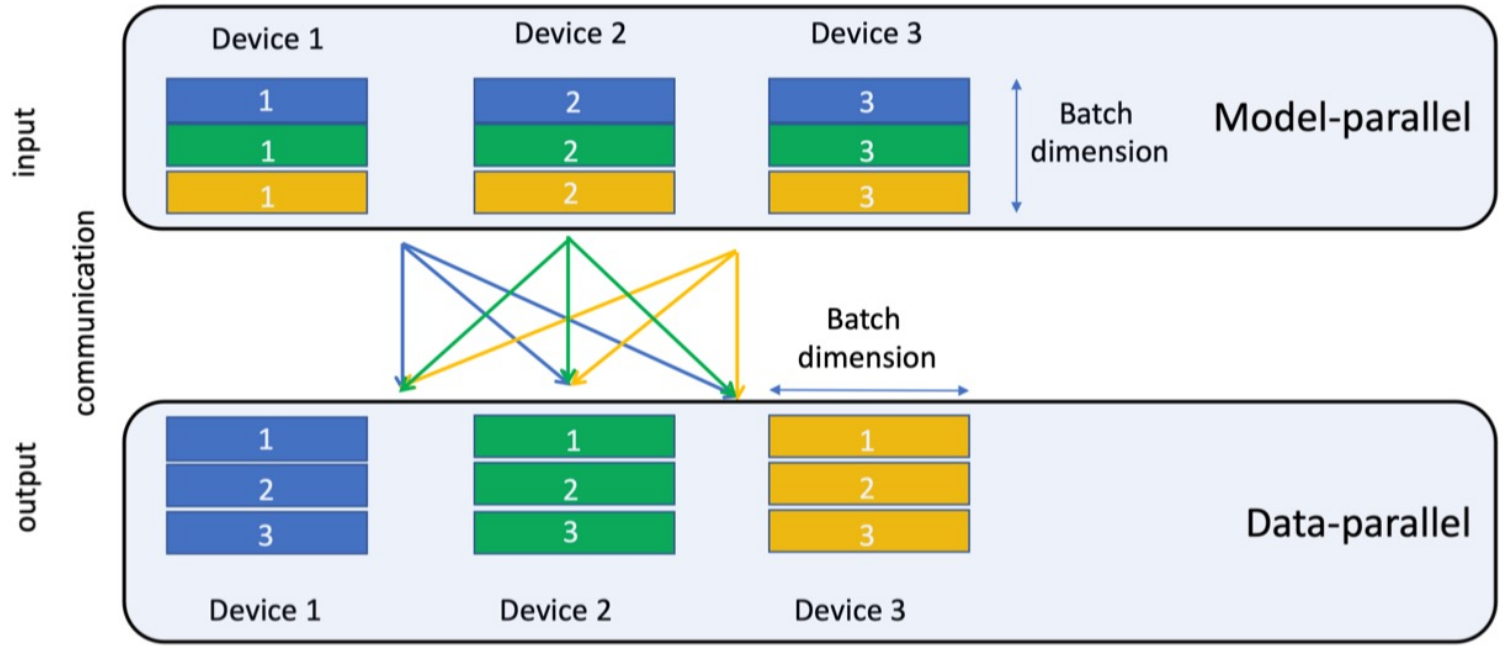
- Wide 部分的主要作用是让模型具有较强的“记忆能力”（Memorization）；
- Deep 部分的主要作用是让模型具有“泛化能力”（Generalization）；



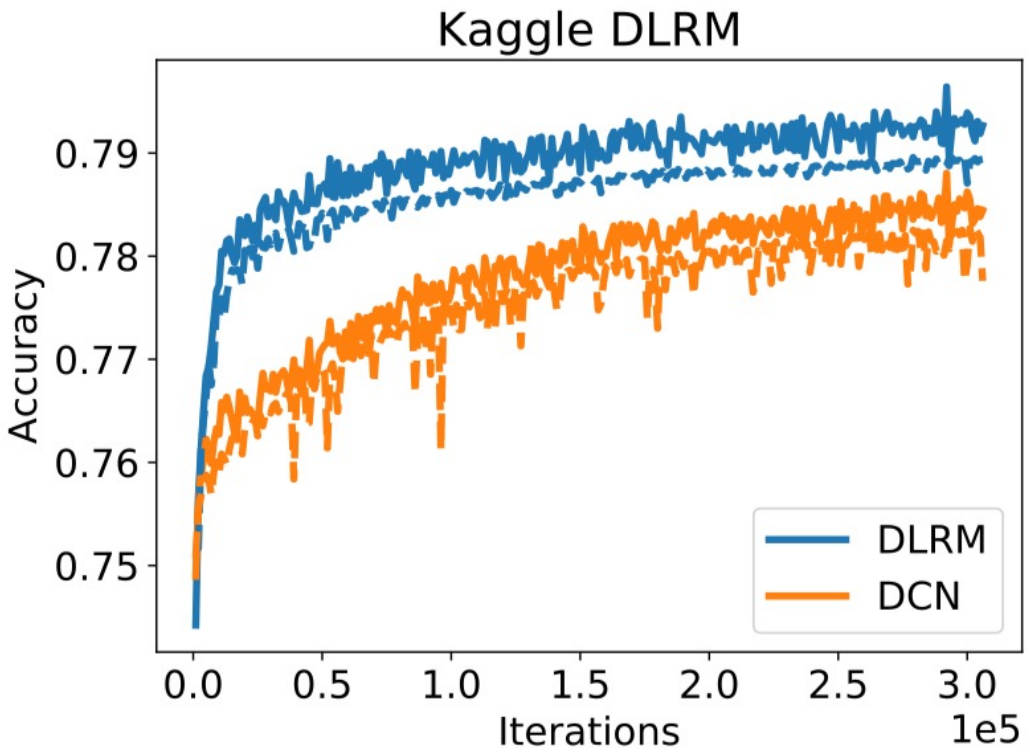
2. DLRM 模型

- 2019~2020 Mate 总结工业界推荐模型设计，高度概括后提出 DLRM（Deep Learning Recommend ation Model）模型架构，针对推荐系统架构的高效和并行问题提出解决方案。

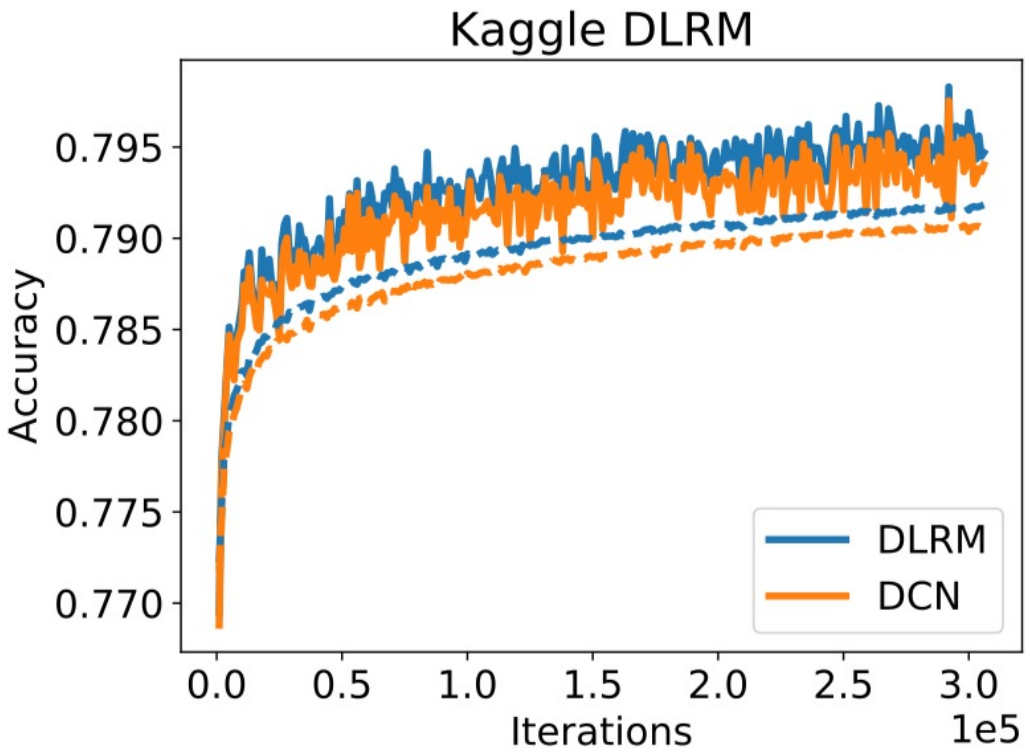




2. DLRM 模型效果



(a) SGD



(b) Adagrad

Summary

1. **Google wide&deep：**
   - 推荐系统的 CTR 估计第一次引入深度学习并极大改变了推荐系统的技术路线；
2. **Meta DLRM：**
   - 总结工业界推荐模型 wide&deep 的变种设计，高度概括后提出 DLRM 架构；

03 瓶颈期

2021~2023

Understanding Capacity-Driven Scale-Out Neural Recommendation Inference
Understanding Scaling Laws for Recommendation Models

1. Scale-Out DLRM

- Meta DLRM 模型规模突破 TB 级别，占用计算中心 79%资源，QPS 时延压力大，急需生产环境中横向扩展 DLRM 分布式推理架构。

- 工业界大规模 TB 级 DRLM 分布式推理架构
- 分析链路 E2E 对时延的影响，大 Embedding 分片架构

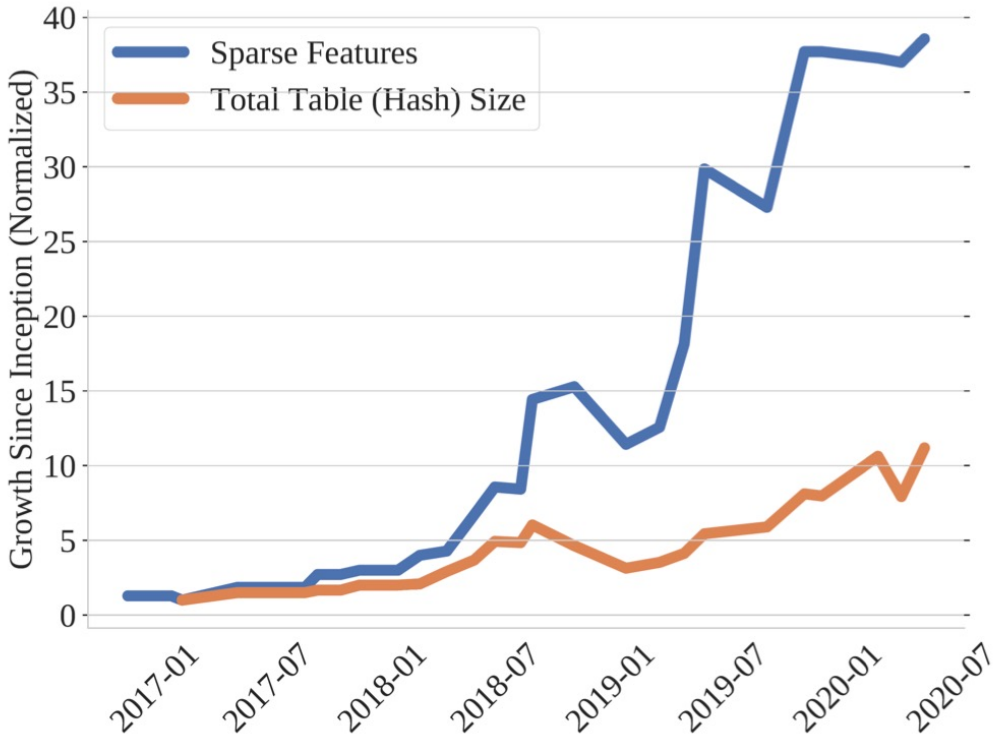
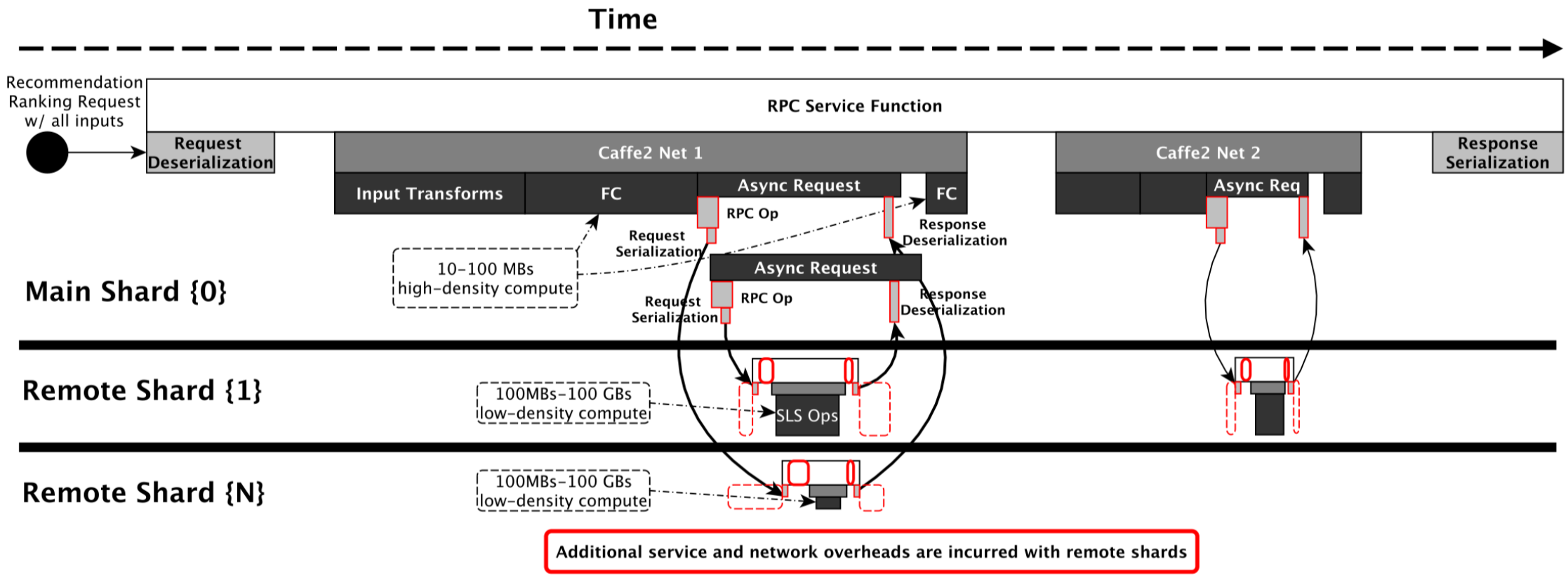


Fig. 1: Historical model growth of significant production recommendation model. Both number of features and embeddings have grown an order of magnitude in only three years.

1. Scale-Out DLRM

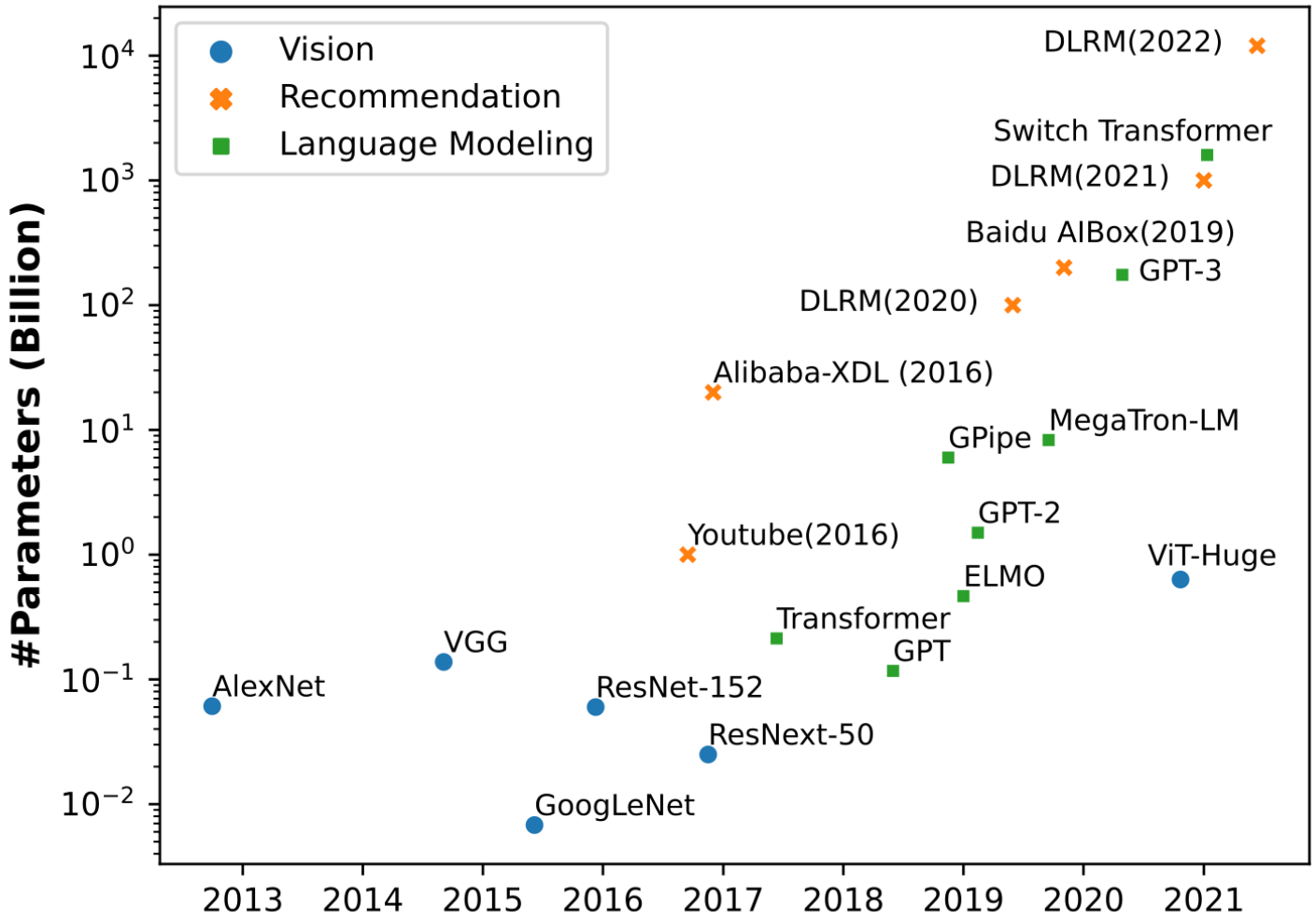
- 提出横向扩展的 Embedding 分片算法和网络通信优化方案



2. Understanding Scaling Laws for Recommendation Models

- 2018~2022 年，工业级推荐模型参数规模增长了 4 个等级。

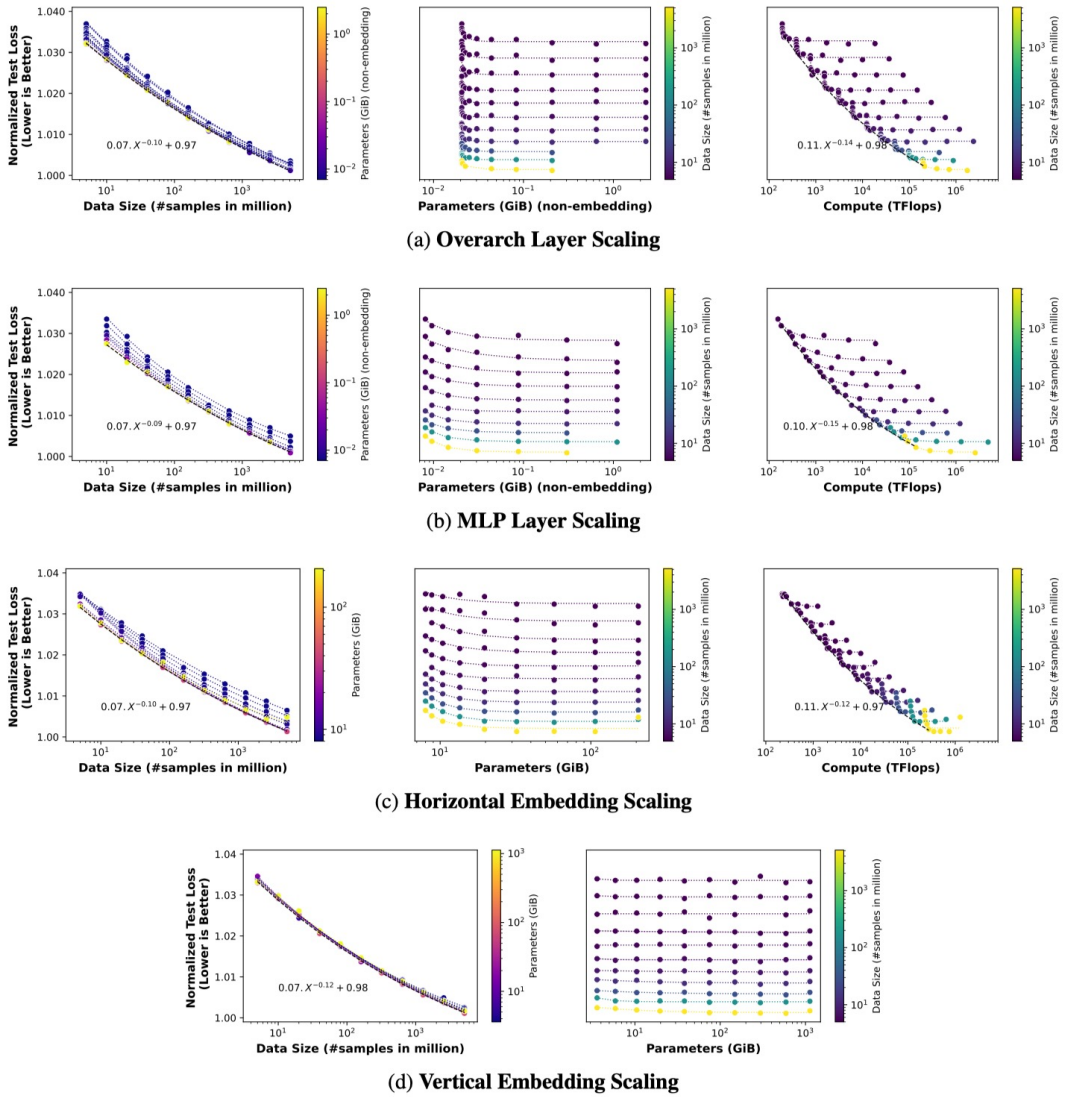
- 文中定量分析模型参数量 N，训练算力 C，训练数据量 D 随着规模增长，对于推荐精度的影响。



2. Understanding Scaling Laws for Recommendation Models

- 在更好的模型架构架构出现前，训练数据量 D 的 Scaling 仍然有少幅度的提升作用。

- 在当前 DLRM 架构下， 推荐模型参数量 N 的 Scaling 已经逼近极限，继续提升精度收益减少。



(a) Overarch Layer Scaling

(b) MLP Layer Scaling

(c) Horizontal Embedding Scaling

(d) Vertical Embedding Scaling

2. Understanding Scaling Laws for Recommendation Models

- **提升训练数据量：**DLRM 架构下模型参数量已经饱和，未来重点可以考虑提升训练数据量。

- **β 参数效果提升：**对新模型结构的探索，增加数据仅仅获得 α 的提升，长期方案应该追求 β 的效果提升。

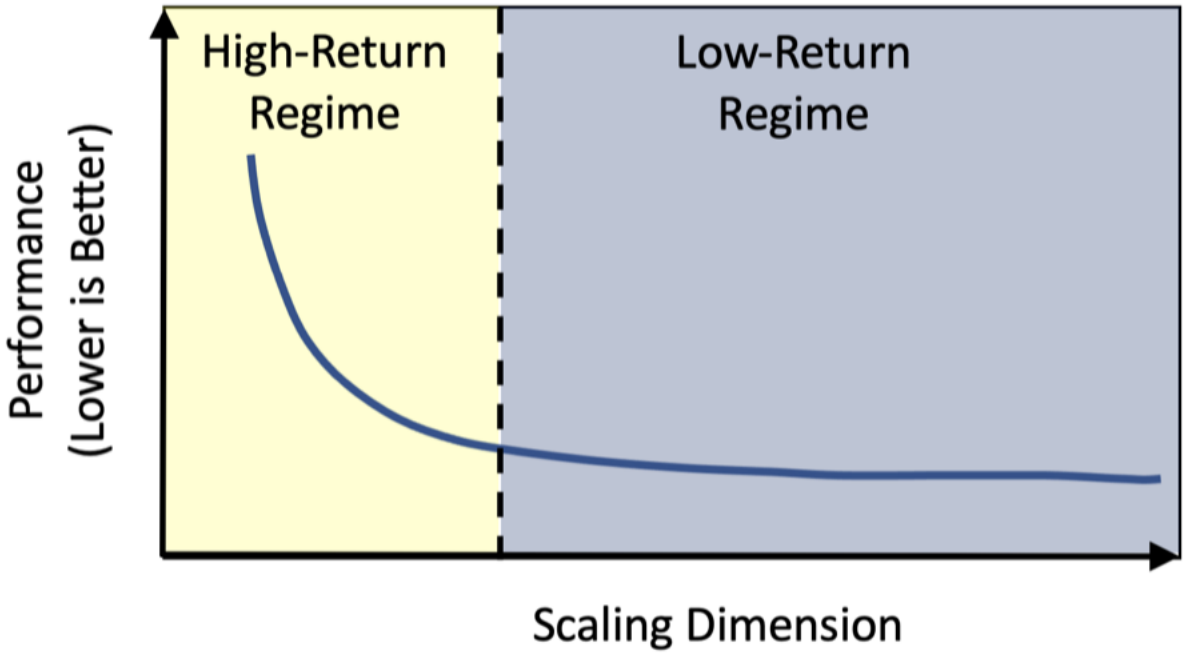


Figure 3: Power-Law Function Characterization.

	α	β	γ	R^2	Sat.	Best Scaling Approach	Ref.
Data Scaling Efficiency	0.07	[0.09 - 0.12]	[0.97 - 0.98]	0.999	No	$V > H > O > M$	Fig 4
Parameter Scaling Efficiency	(0 - 0.5]	[0.4 - 7.6]	[0.97 - 1]	0.9	Yes	$V \approx H \approx O \approx M$	Fig 6
Compute Scaling Efficiency	0.11	[0.12-0.15]	0.98	0.999	No	$M > O > H$	Fig 5

Summary

1. **Scale-Out DLRM：**
   - TB 级别规模 DLRM 在线推理成本、时延增加，提出分布式推理架构；
2. **Scaling Laws DLRM：**
   - DLRM 架构下 TB 级别模型精度基本饱和，继续提升依赖于新架构的出现；

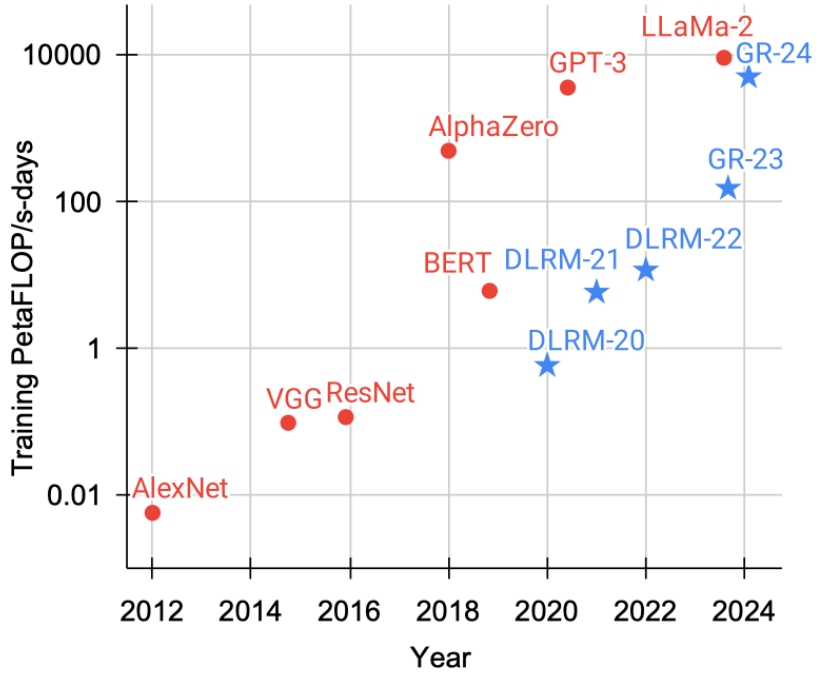
03 新范式

2023~2024

使用 Transformer 作为推荐可能性

- 采用 Transformer 架构作为推荐：

  1. **长序列支持：** LLM Google 发表支持 10M 词表，与推荐十亿规模词表有能够相比较；
  2. **有序生成：**生成下一个 Tokens 可以作为推荐排序内容，LLM Output 解决了排序的问题；



使用 Transformer 作为推荐难点

- 直接采用 Transformer 架构的挑战：
  1. **特征缺乏显式结构性**：海量异构特征，如高基数 IDs、交叉特征、物品和用户特征等；
  2. **十亿级别动态词表**：语言模型 100K 词表 vs 推荐十亿规模词表，且实时调整；
  3. **生成式计算开销大**：大规模推荐吞吐和访问量巨大，交互行为数据规模远超语言大模型；

Actions Speak Louder than Words

- 生成式推荐（Generative Recommenders）新范式，将推荐系统的主要任务（召回+排序）转化为生成模型框架内的序列任务。
- 1.5 万亿参数的生成式推荐模型，不仅在在线A/B测试中取得了12.4%的性能提升，而且已经被部署在 Meta 服务于数十亿用户的互联网平台。

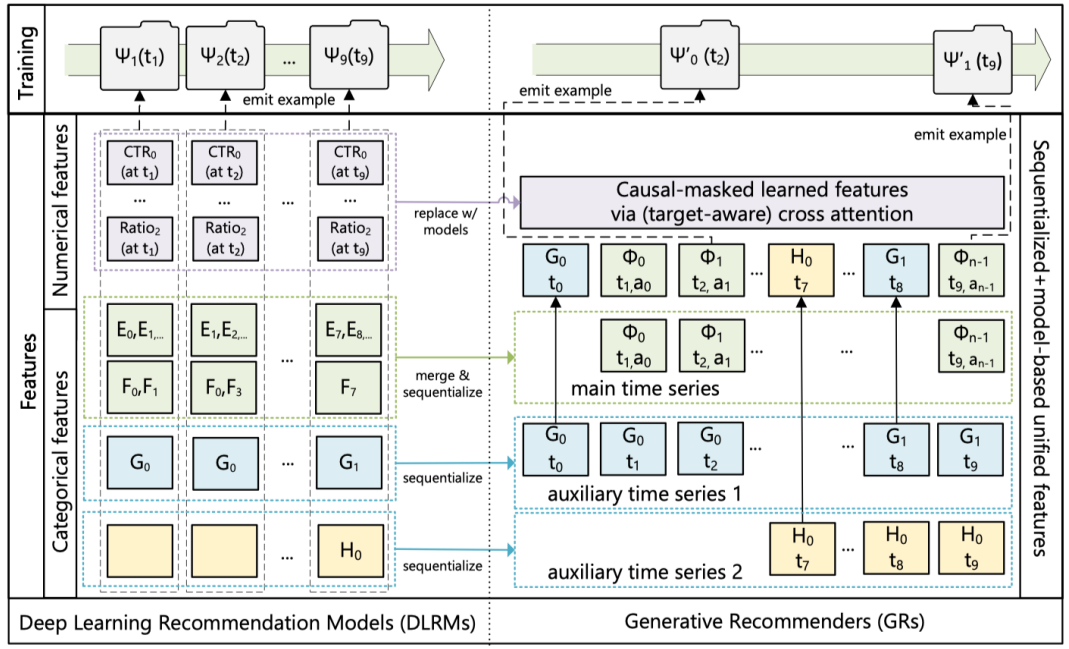
Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations

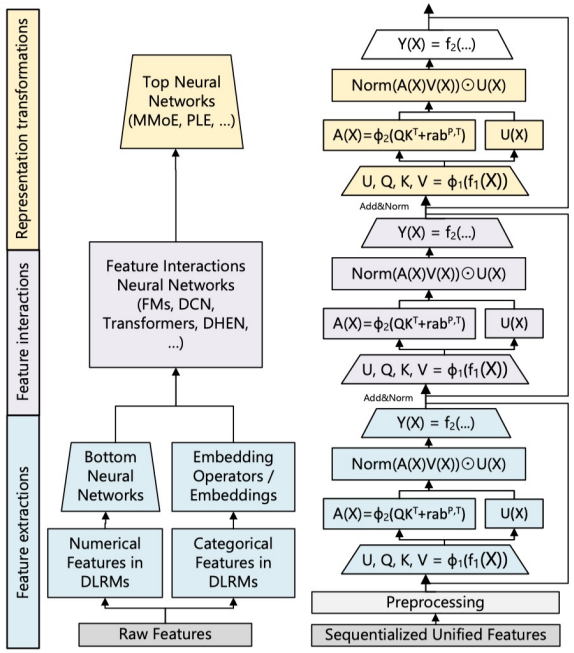
---

Jiaqi Zhai [1] **Lucy Liao** [1] **Xing Liu** [1] **Yueming Wang** [1] **Rui Li** [1]
Xuan Cao [1] **Leon Gao** [1] **Zhaojie Gong** [1] **Fangda Gu** [1] **Michael He** [1] **Yinghai Lu** [1] **Yu Shi** [1]

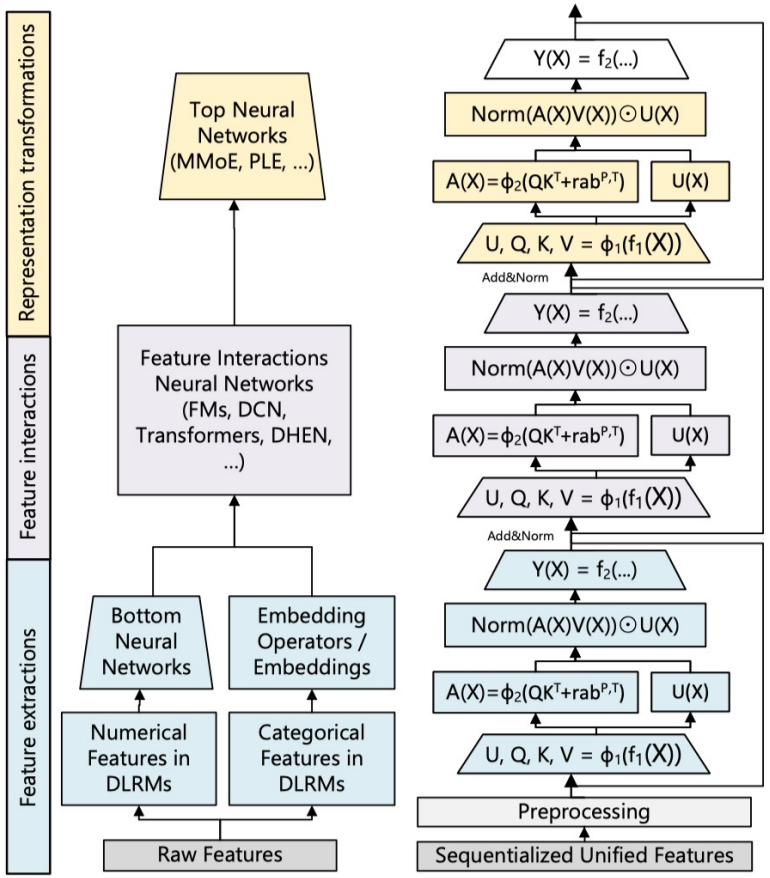
Actions Speak Louder than Words

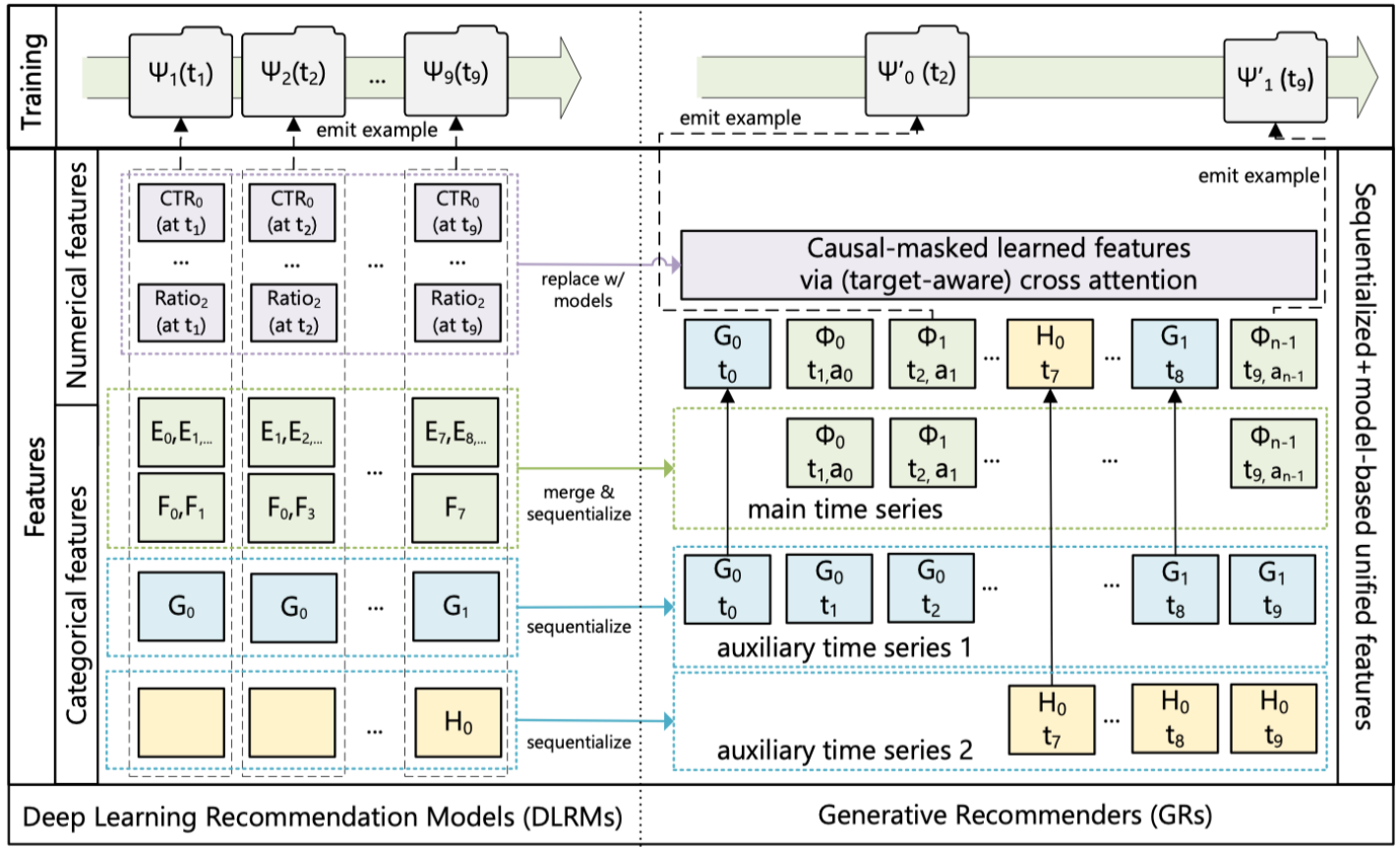
- **生成式建模：**将找回和排序定义为序列直推式生成任务，Transformer 结构实现特征自动交互
- **HSTU Encoder 架构：**Pointwise Attention 代替 Softmax Attention，适配大规模动态词表
- **性能优化：**流式训练采样、序列随机长度、推理小 Batch 并行加速 M-Falcon 算法





Actions Speak Louder than Words





04 小节与思考

总结

1. 召回排序、召回排序烦烦烦，E2E 大模型统一一切！统一自动驾驶，统一生成式推荐！
2. Google 引入 DLRM、Meta 实践优化，Scaling 后进入瓶颈期后迎来大模型继续 Scaling Law！

未来互联网厂商对于推荐的发展

1. LLM 训练业务成熟稳定，并掌握训练技巧后，互联网部门头部厂商将会慢慢转向生成式推荐，有望成为算力填充的下一个消耗资源大头。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/AIFoundation

Reference

1. Naumov M, Mudigere D, Shi H J M, et al. Deep learning recommendation model for personalization and recommendation systems[J]. arXiv preprint arXiv:1906.00091, 2019.
2. DLRM: An advanced, open source deep learning recommendation model
3. Wide & Deep Learning for Recommender Systems