



DeepSeek系列介绍

张 禹

2025.02.27



本次汇报的目标

- 关注DeepSeek的内在核心技术
- 按照时间线介绍DeepSeek如何一步步发展至今
- 力求不同研究背景的同学都能听懂
- 不关注技术上的细枝末节

- DeepSeek是什么？

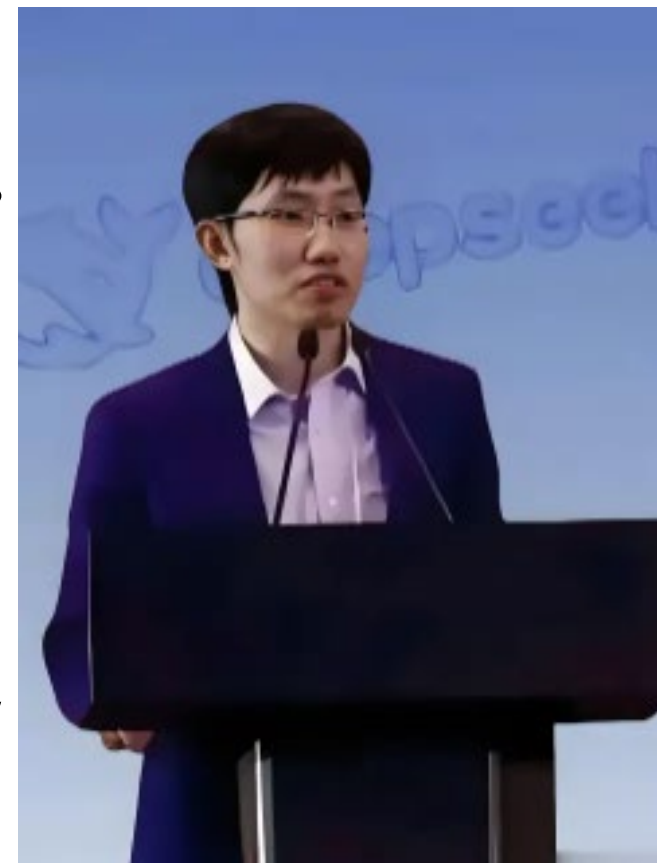
Deepseek是杭州深度求索人工智能技术研究中心，也是该公司开发的多模态大语言模型（MLLM）

- DeepSeek掌舵人——梁文峰

梁文峰，中国人工智能公司DeepSeek的创始人兼首席执行官。1985年生，于2007年获得浙江大学电子信息工程学士学位，2010年取得信息与通信工程硕士学位，师从项志宇教授，主要研究机器视觉领域。

在攻读硕士期间，梁文峰与同学组建团队，收集金融市场数据，探索将机器学习应用于全自动量化交易。2013年，他创立了杭州雅克比投资管理有限公司，将人工智能与量化交易相结合。2015年，他们进一步创立了杭州幻方科技有限公司，并于2016年成立了宁波幻方量化投资管理合伙企业，专注于利用数学和人工智能进行量化投资。

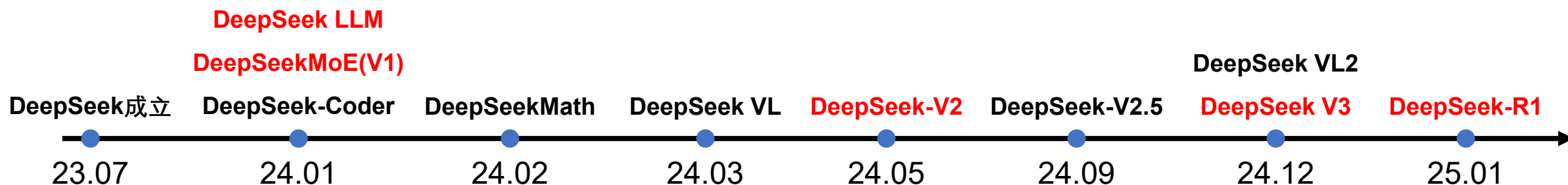
在量化投资领域取得成功，梁文峰将目光投向人工智能的更广阔应用。2023年7月，他宣布进军通用人工智能（AGI）领域，创立了DeepSeek，专注于大型语言模型的研究与开发。DeepSeek迅速崛起，其发布的DeepSeek-V3和DeepSeek-R1模型在性能和成本方面表现出色，引起全球关注。



● DeepSeek为什么爆火?

1. 模型性能好，体验佳
2. 开源的多模态大语言模型
3. 训练成本低 (R1 560万美元)
4. 中国的模型

● DeepSeek系列发展时间线



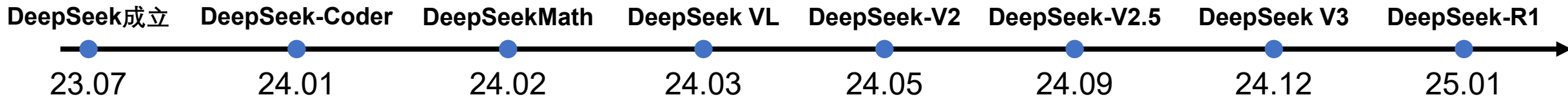


DeepSeek LLM: Scaling Open-Source Language Models with Longtermism

DeepSeek LLM

DeepSeekMoE(V1)

DeepSeek VL2



- 该工作的动机是什么？

想要进一步深入探究LLM中的Scaling Law：前人在LLM的Scaling Law上做过一些探索，但是这些探索依然存在很大的不确定性，已有的很多结论、观点都是相矛盾的

- Scaling Law是什么？

Scaling Law（缩放定律）是人工智能领域，尤其是大型语言模型研究中的一个重要概念。于2020年首次被OpenAI团队提出，它描述了**模型性能**（如准确率、生成质量）与**模型规模**（参数数量）、**数据集大小**以及**计算资源**投入之间的数学关系。

- ◆ 模型规模 (N, 参数量)：指的是模型中的神经网络参数总数
- ◆ 数据集大小 (D, 数据量)：指的是训练模型时使用的token(词或子词)数量
- ◆ 计算资源 (C, 算力)：指的是训练模型时消耗的计算量，通常以浮点运算次数(FLOPs)衡量

研究发现，模型的性能（通常用损失函数值来衡量）与这三个变量之间存在幂律关系

研究发现，模型的性能（通常用损失函数值来衡量）与这三个变量之间存在幂律关系，在数学上，这种关系常被表示为：

$$L \propto N^{-\alpha} \cdot D^{-\beta} \cdot C^{-\gamma}$$

其中：

- L 是模型损失（越小越好）
- α, β, γ 是幂律函数，具体值取决于任务和模型结构

假设有一个1亿参数量语言模型，用10亿token的数据训练，性能达到某水平。如果按照Scaling Law：

- 把模型参数增加到10亿（10倍），性能可能提升20%
- 把数据量增加到100亿tokens（10倍），性能可能再提升15%
- 把算力翻倍，性能可能再小幅提升

● 模型与数据

改进1：数据集和数据处理方面。数据集更大了（2万亿tokens），采用了**去重**、**过滤**、**混合**三种方法，使得数据集的丰富性和多样性很大的提升。数据处理层面，原来采用了字符对编码Byte Pair Encoding (BPE)，现在采用了字节层面的字符对编码Byte-Level Byte Pair Encoding (BBPE)

Byte Pair Encoding (BPE): low lower newest widest

- ◆ 初始化：将每个单词分解为字符序列：l o w l o w e r n e w e s t w i d e s t
- ◆ 统计频率 (计算所有相邻字符对的频率)：l和o: 2 o和w: 2 w和e: 2 e和s: 2 s和t: 2 其他为1
- ◆ 合并频率最高的字节对：合并l o为lo，更新序列：lo w lo w e r n e w e s t w i d e s t
- ◆ 重复合并：继续合并频率最高的字节对，直到达到预定的词汇表大小
- ◆ 生成词汇表：词汇表可能包括：l, o, w, e, r, n, s, t, lo, we

● 模型与数据

改进1：数据集和数据处理方面。数据集更大了（2万亿tokens），采用了**去重**、**过滤**、**混合**三种方法，使得数据集的丰富性和多样性很大的提升。数据处理层面，原来采用了字符对编码Byte Pair Encoding (BPE)，现在采用了字节层面的字符对编码Byte-Level Byte Pair Encoding (BBPE)

Byte-Level Byte Pair Encoding (BBPE): café naïve

◆ 初始化：将每个单词转换为字节序列（假设使用UTF-8）

café -> [99, 97, 102, 195, 169]

naïve -> [110, 97, 195, 175, 118, 101]

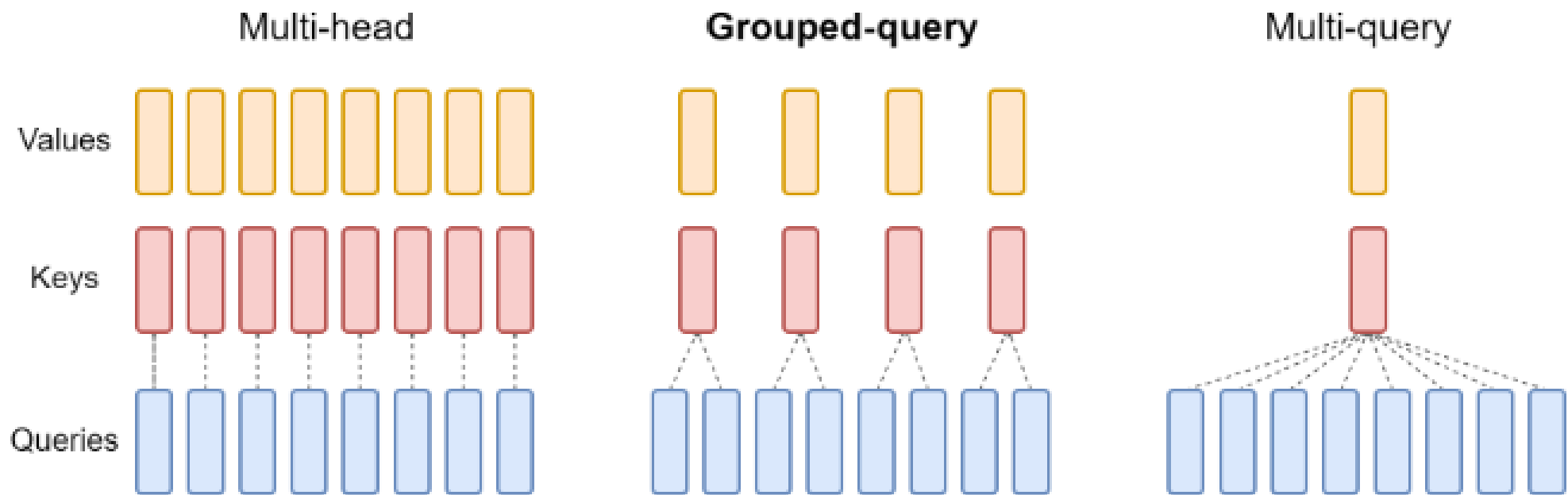
◆ 统计频率 (计算所有相邻字节对的频率): [99, 97]: 1 [97, 102]: 1 [102, 195]: 1 [195, 169]: 1
[110, 97]: 1 [97, 195]: 1 [195, 175]: 1 [175, 118]: 1 [118, 101]: 1

◆ 合并与重复合并

◆ 生成词汇表：词汇表可能包括: c, af, é, na, ï, ve

● 模型与数据

改进2：模型结构。 DeepSeek LLM 在微观设计上主要遵循 LLaMA，有7B（30层）和67B（95层）两个版本。采用了分组查询注意力机制 **Grouped-Query Attention (GQA)**。



[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[2] Shazeer, Noam. "Fast transformer decoding: One write-head is all you need." arXiv preprint arXiv:1911.02150 (2019).

[3] Ainslie, Joshua, et al. "Gqa: Training generalized multi-query transformer models from multi-head checkpoints." arXiv preprint arXiv:2305.13245 (2023).

- **模型与数据**

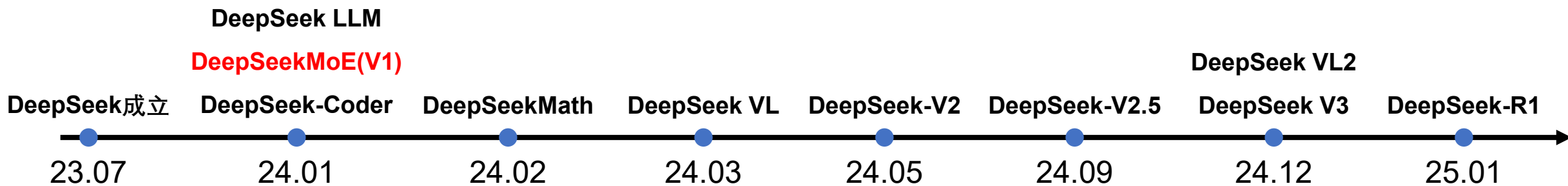
改进3：训练方式的改进和对齐方式的改进：采用监督微调（SFT）和直接偏好优化（DPO）

- **Scaling Law**

- ◆ 建立了超参数的缩放定律，为确定最佳超参数提供了一个经验框架。
- ◆ 采用非嵌入FLOPs/token-M来表示模型规模，更精确的最佳模型/数据扩展分配策略和大规模模型的泛化损失的更好预测。
- ◆ 提出了一个基于计算预算和数据规模的模型扩展策略，该策略在计算预算有限的情况下提高了模型的泛化能力。
- ◆ 证明了缩放定律的普遍适用性，并预测了未来更大规模模型的性能。



DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models

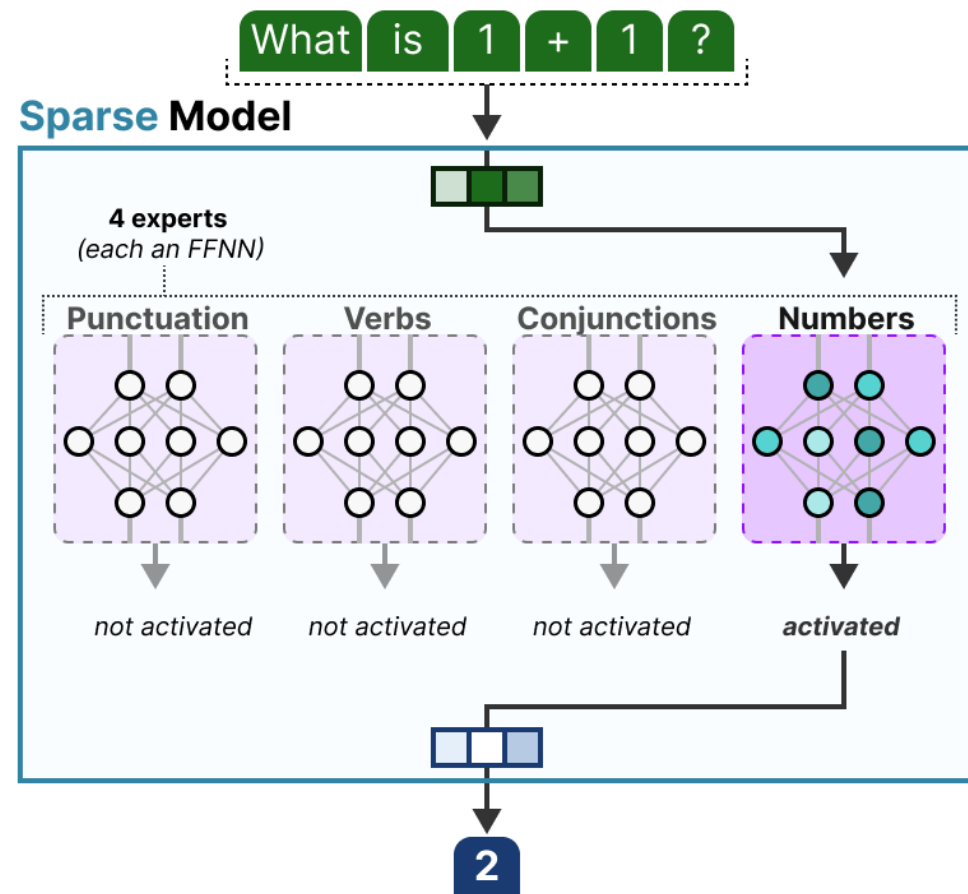


● 该工作的动机是什么？

DeepSeek LLM的探究指出了在足够的多的、足够好的训练数据下，可以通过增加参数量和计算量来扩大语言模型的规模得到更大的计算模型。然而其计算成本是极高的，如何降本增效？

● MoE是什么？

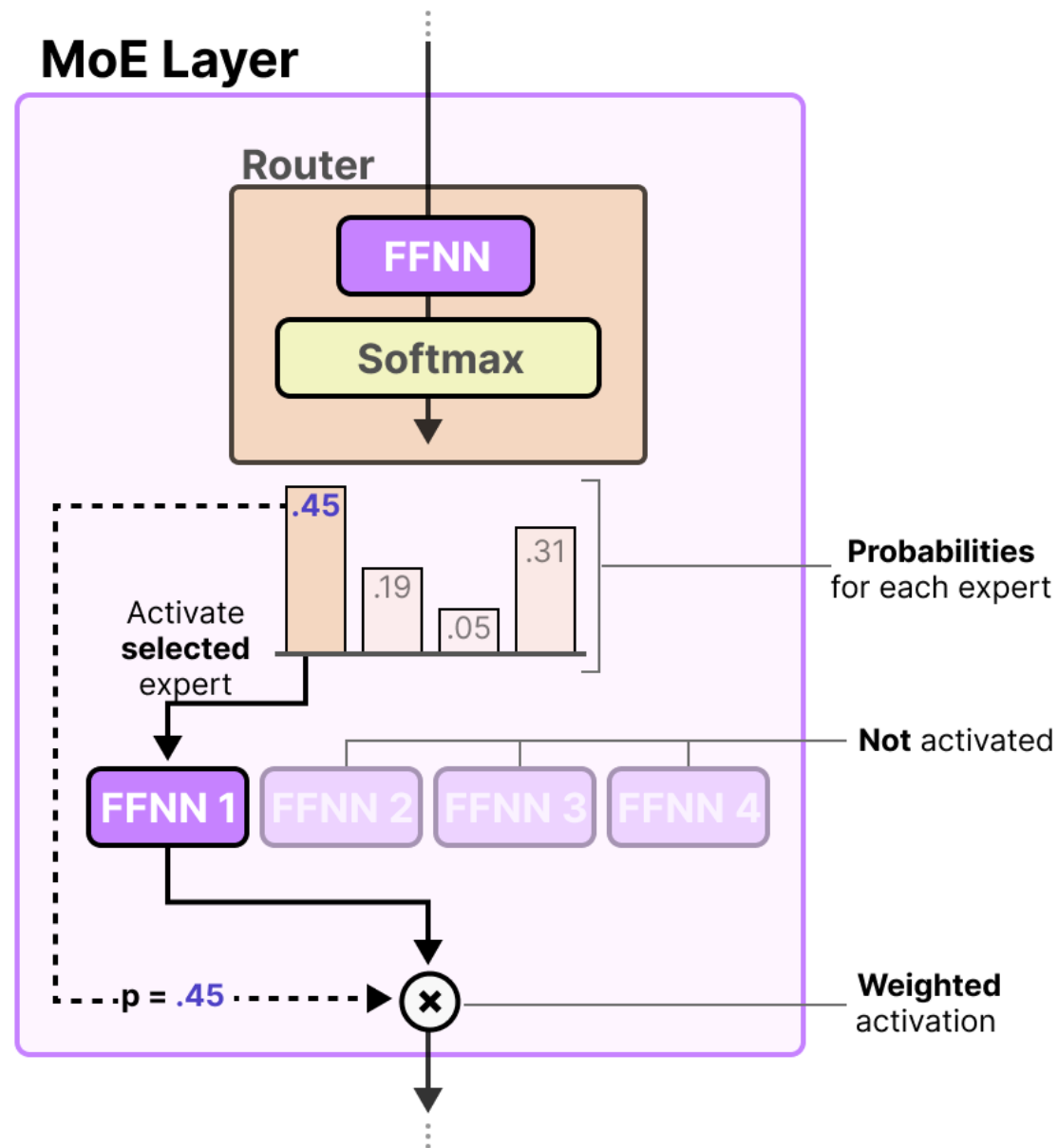
混合专家模型（Mixture of Experts, MoE）是一种先进的神经网络架构。旨在通过整合多个模型或“专家”的预测来提升整体模型性能。MoE模型的核心思想是将输入数据分配给不同的专家子模型，然后将所有子模型的输出进行合并，以生成最终结果。这种分配可以根据输入数据的特征进行动态调整，确保每个专家处理其最擅长的数据类型或任务方面，从而实现更高效、准确的预测。



DeepSeekMoE (V1)



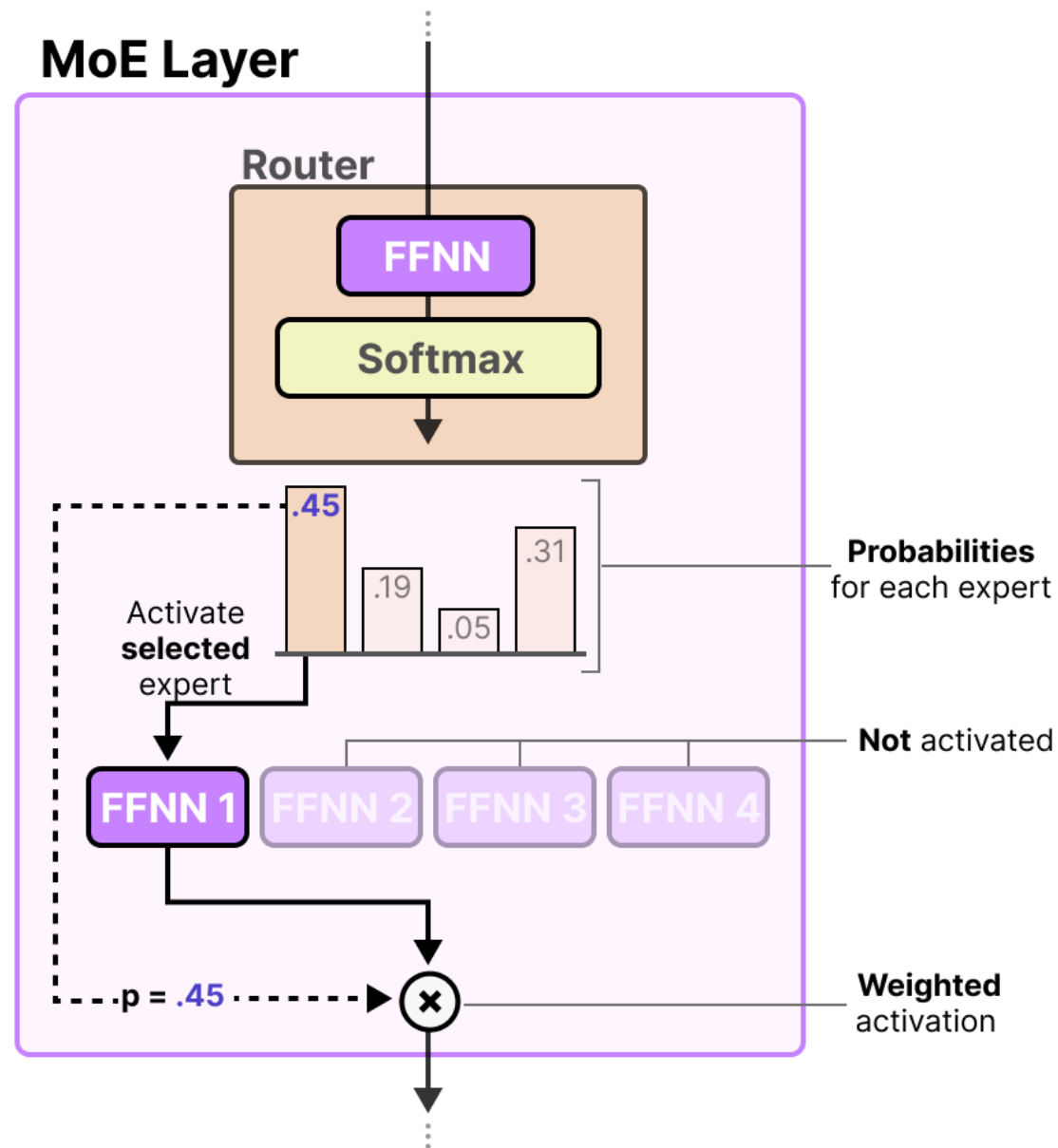
- ◆ **专家 (Experts):** 模型中的每个专家都是一个独立的神经网络，专门处理输入数据的特定子集或特定任务。例如，在自然语言处理任务中，一个专家可能专注于处理与语言语法相关的内容，而另一个专家可能专注于语义理解。
- ◆ **门控网络 (Gating Network) :** 门控网络的作用是决定每个输入样本应该由哪个专家或哪些专家来处理。它根据输入样本的特征计算出每个专家的权重或重要性，然后根据这些权重将输入样本分配给相应的专家。门控网络通常是一个简单的神经网络，其输出经过softmax激活函数处理，以确保所有专家的权重之和为1。



DeepSeekMoE (V1)



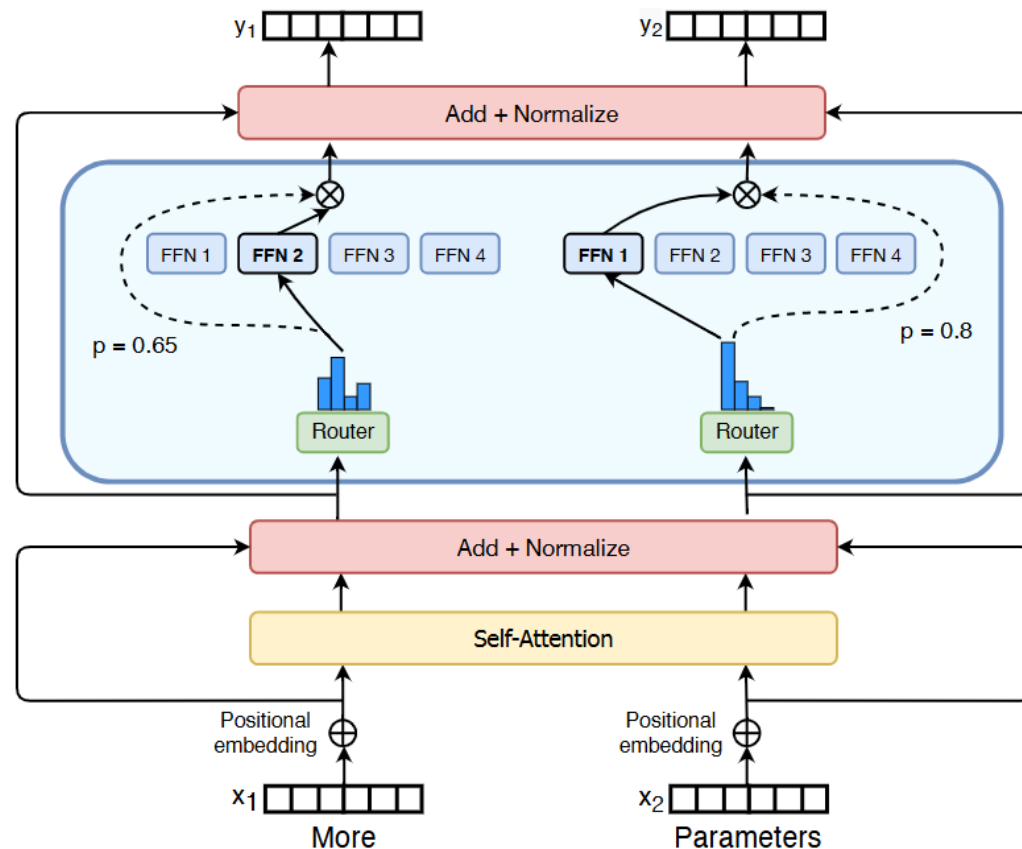
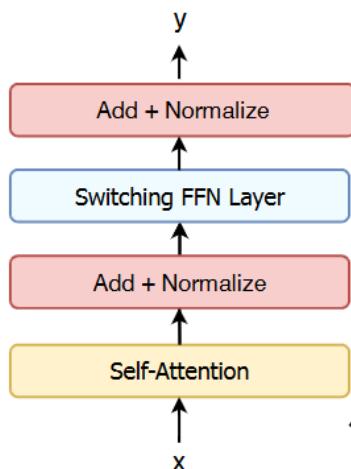
- ◆ **训练与优化：**专家的训练过程与传统的神经网络类似，通过反向传播算法根据损失函数的梯度更新网络参数。由于router还负责分配与权重，因此需要与router一起协同训练和优化。
- ◆ **Top-K选择策略：**除了根据softmax权重进行加权求和外，router还可以采用Top-K选择策略来进一步简化计算和提高效率。
- ◆ **负载均衡与辅助损失函数：**为了确保各个专家之间的负载均衡，避免某些专家过度负担而其他专家闲置的情况，通常会在MoE模型的训练过程中引入辅助损失函数。辅助损失函数的目标是使每个专家处理的输入数据量大致相等，从而提高模型的整体效率和稳定性。



DeepSeekMoE (V1)



● MoE + Transformer



[1] Jacobs, Robert A., et al. "Adaptive mixtures of local experts." Neural computation 3.1 (1991): 79-87.

[2] Lepikhin, Dmitry, et al. "Gshard: Scaling giant models with conditional computation and automatic sharding." arXiv preprint arXiv:2006.16668 (2020).

[3] Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." Journal of Machine Learning Research 23.120 (2022): 1-39.

● 细粒度专家策略

在典型 MoE 架构之上，将每个专家 FFN 细分为m个较小的专家

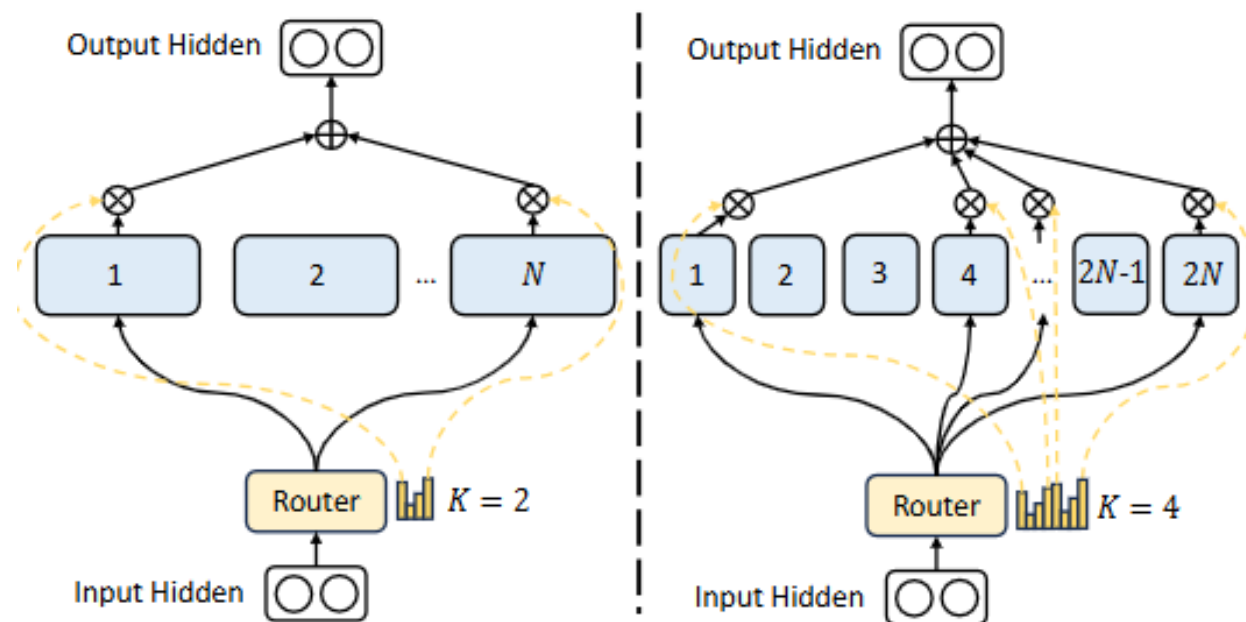
$$\mathbf{u}_{1:T}^l = \text{Self-Att}(\mathbf{h}_{1:T}^{l-1}) + \mathbf{h}_{1:T}^{l-1},$$

$$\mathbf{h}_t^l = \text{FFN}(\mathbf{u}_t^l) + \mathbf{u}_t^l,$$

$$\mathbf{h}_t^l = \sum_{i=1}^N \left(g_{i,t} \text{FFN}_i(\mathbf{u}_t^l) \right) + \mathbf{u}_t^l,$$

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{Topk}(\{s_{j,t} | 1 \leq j \leq N\}, K), \\ 0, & \text{otherwise,} \end{cases}$$

$$s_{i,t} = \text{Softmax}_i(\mathbf{u}_t^{lT} \mathbf{e}_i^l),$$



(a) Conventional Top-2 Routing ➡ (b) + Fine-grained Expert Segmentation ■

● 细粒度专家策略

在典型 MoE 架构之上，将每个专家 FFN 细分为m个较小的专家

$$\mathbf{h}_t^l = \sum_{i=1}^N \left(g_{i,t} \text{FFN}_i(\mathbf{u}_t^l) \right) + \mathbf{u}_t^l,$$

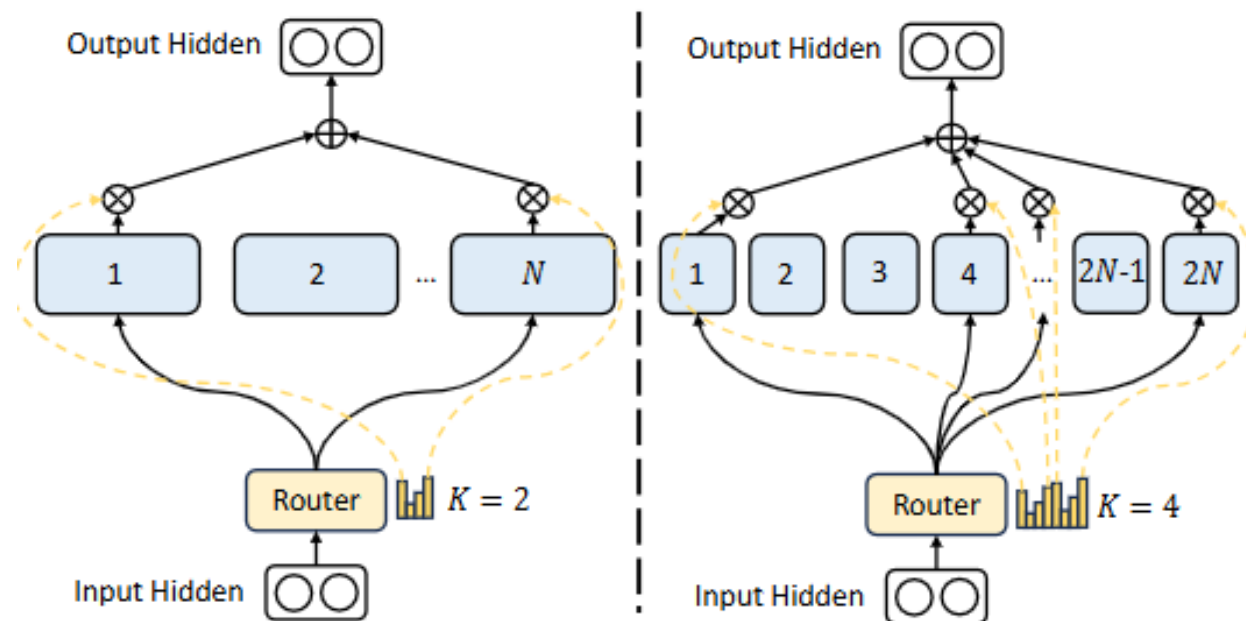
$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{Topk}(\{s_{j,t} | 1 \leq j \leq N\}, K), \\ 0, & \text{otherwise,} \end{cases}$$

$$s_{i,t} = \text{Softmax}_i(\mathbf{u}_t^{lT} \mathbf{e}_i^l),$$

$$\mathbf{h}_t^l = \sum_{i=1}^{mN} \left(g_{i,t} \text{FFN}_i(\mathbf{u}_t^l) \right) + \mathbf{u}_t^l,$$

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{Topk}(\{s_{j,t} | 1 \leq j \leq mN\}, mK), \\ 0, & \text{otherwise,} \end{cases}$$

$$s_{i,t} = \text{Softmax}_i(\mathbf{u}_t^{lT} \mathbf{e}_i^l),$$



(a) Conventional Top-2 Routing \longrightarrow (b) + Fine-grained Expert Segmentation ■

● 共享专家隔离

隔离部分专家以作为共享专家。无论router模块如何，每个token都将确定性地分配给这些共享专家。

$$\mathbf{h}_t^l = \sum_{i=1}^{mN} \left(g_{i,t} \text{FFN}_i \left(\mathbf{u}_t^l \right) \right) + \mathbf{u}_t^l,$$

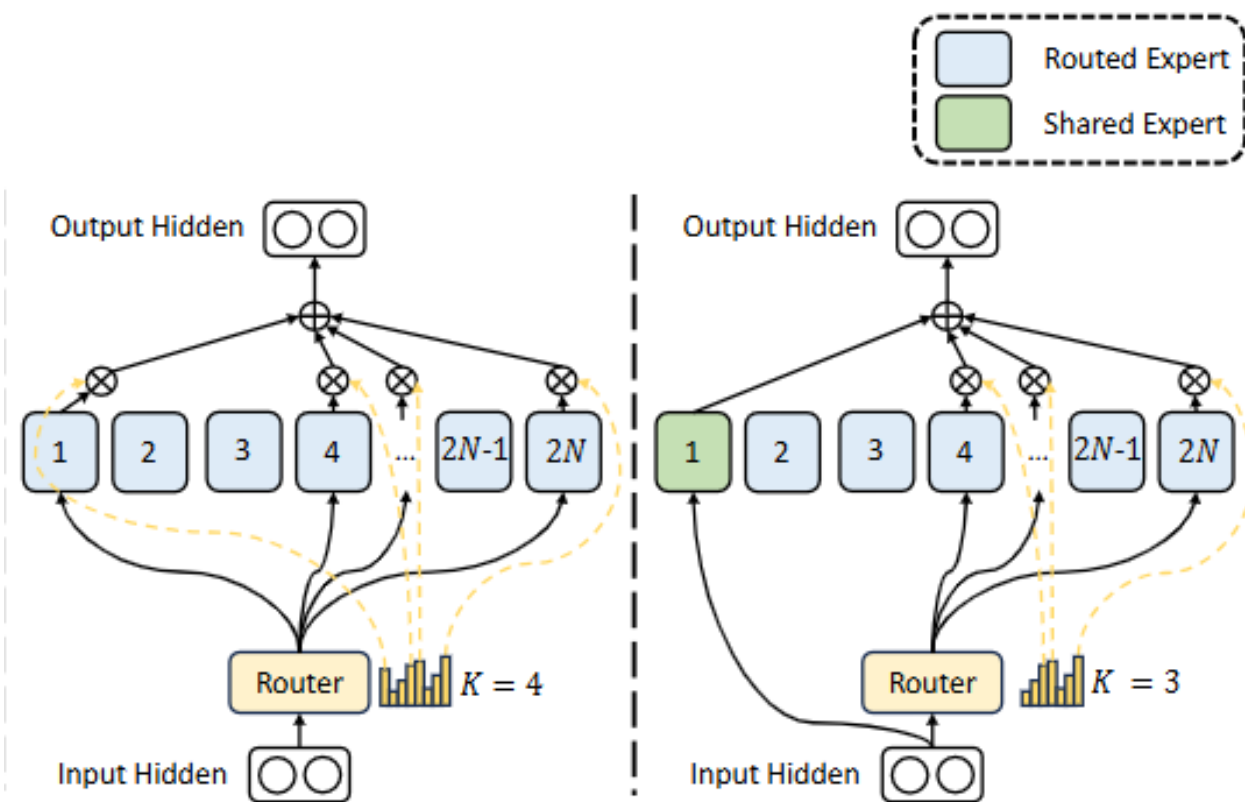
$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{Topk}(\{s_{j,t} | 1 \leq j \leq mN\}, mK), \\ 0, & \text{otherwise,} \end{cases}$$

$$s_{i,t} = \text{Softmax}_i \left(\mathbf{u}_t^{lT} \mathbf{e}_i^l \right),$$

$$\mathbf{h}_t^l = \sum_{i=1}^{K_s} \text{FFN}_i \left(\mathbf{u}_t^l \right) + \sum_{i=K_s+1}^{mN} \left(g_{i,t} \text{FFN}_i \left(\mathbf{u}_t^l \right) \right) + \mathbf{u}_t^l,$$

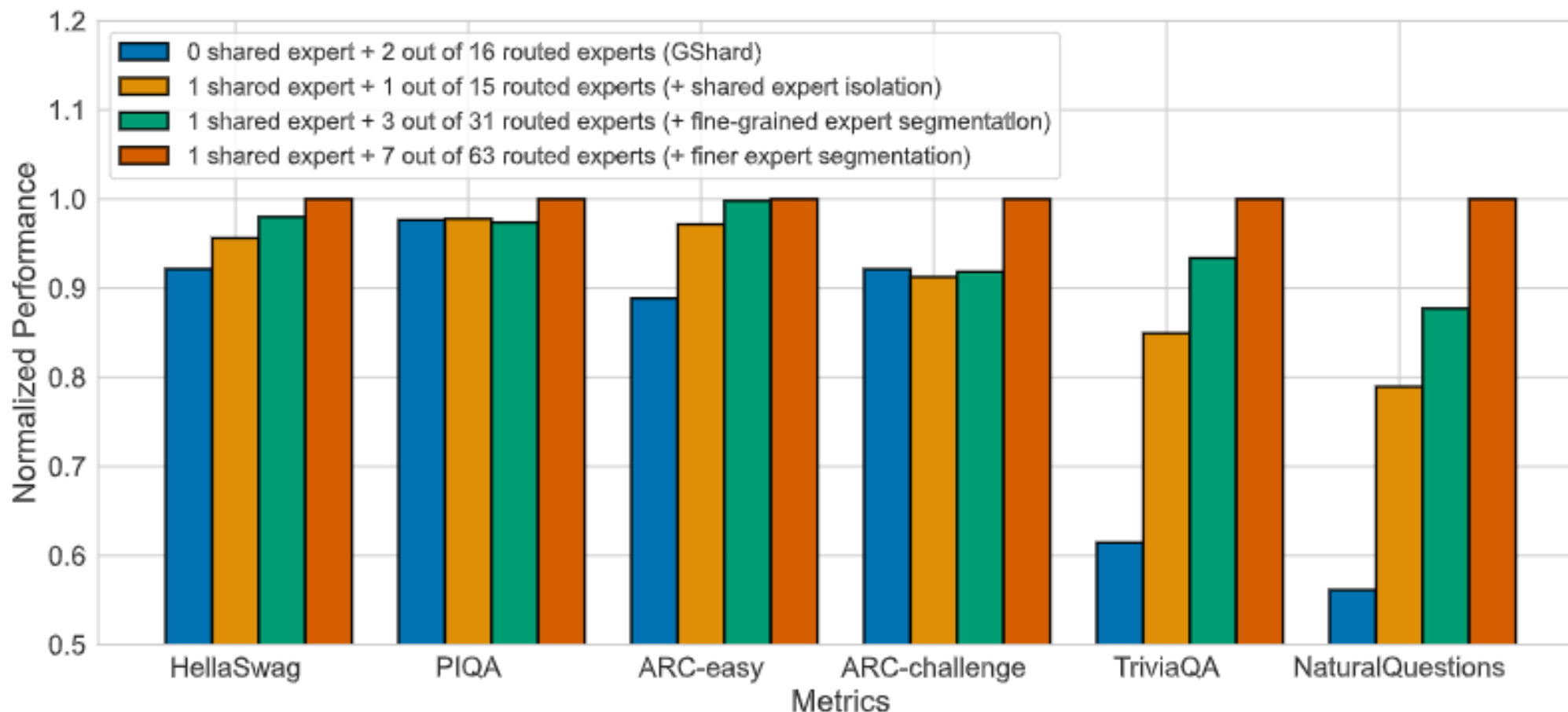
$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{Topk}(\{s_{j,t} | K_s + 1 \leq j \leq mN\}, mK - K_s), \\ 0, & \text{otherwise,} \end{cases}$$

$$s_{i,t} = \text{Softmax}_i \left(\mathbf{u}_t^{lT} \mathbf{e}_i^l \right).$$



(b) + Fine-grained Expert Segmentation \longrightarrow (c) + Shared Expert Isolation (DeepSeekMoE)

DeepSeekMoE (V1)

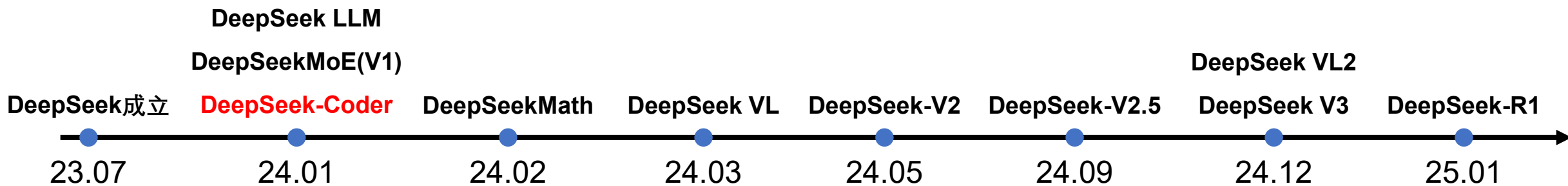


● 负载均衡考虑

专家级平衡损失+设备级平衡损失

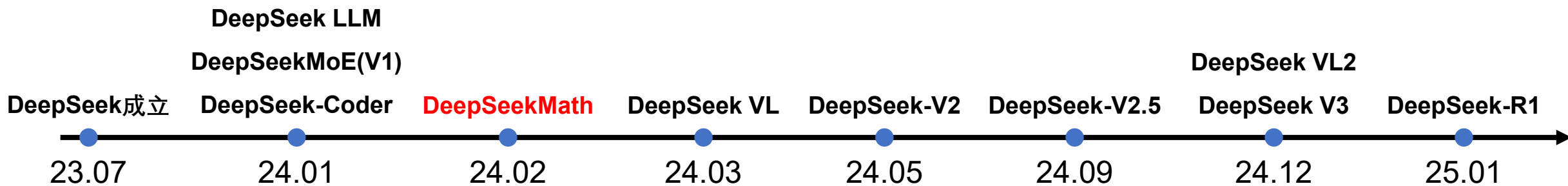


DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence



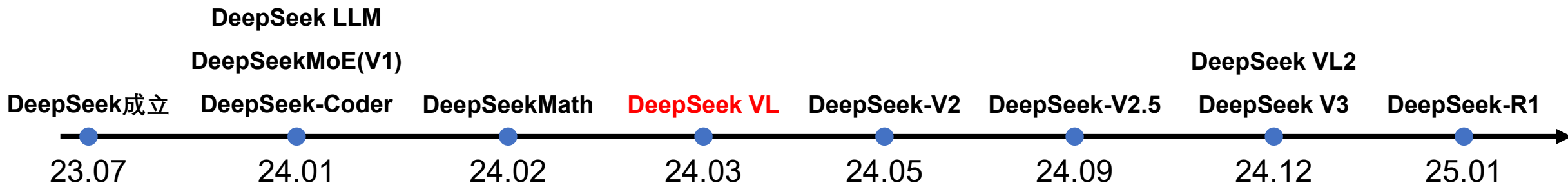


DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models



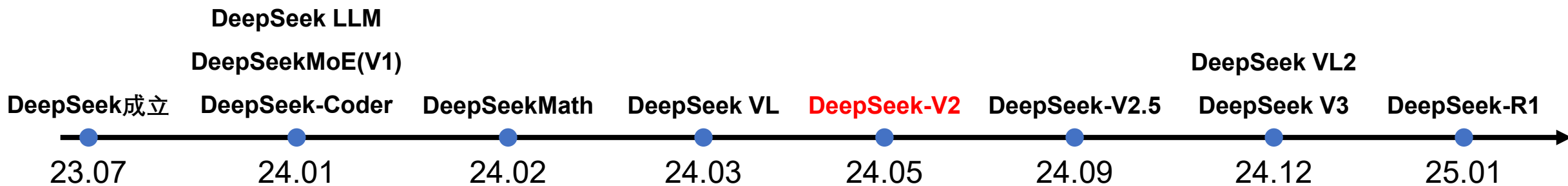


DeepSeek-VL: Towards Real-World Vision-Language Understanding

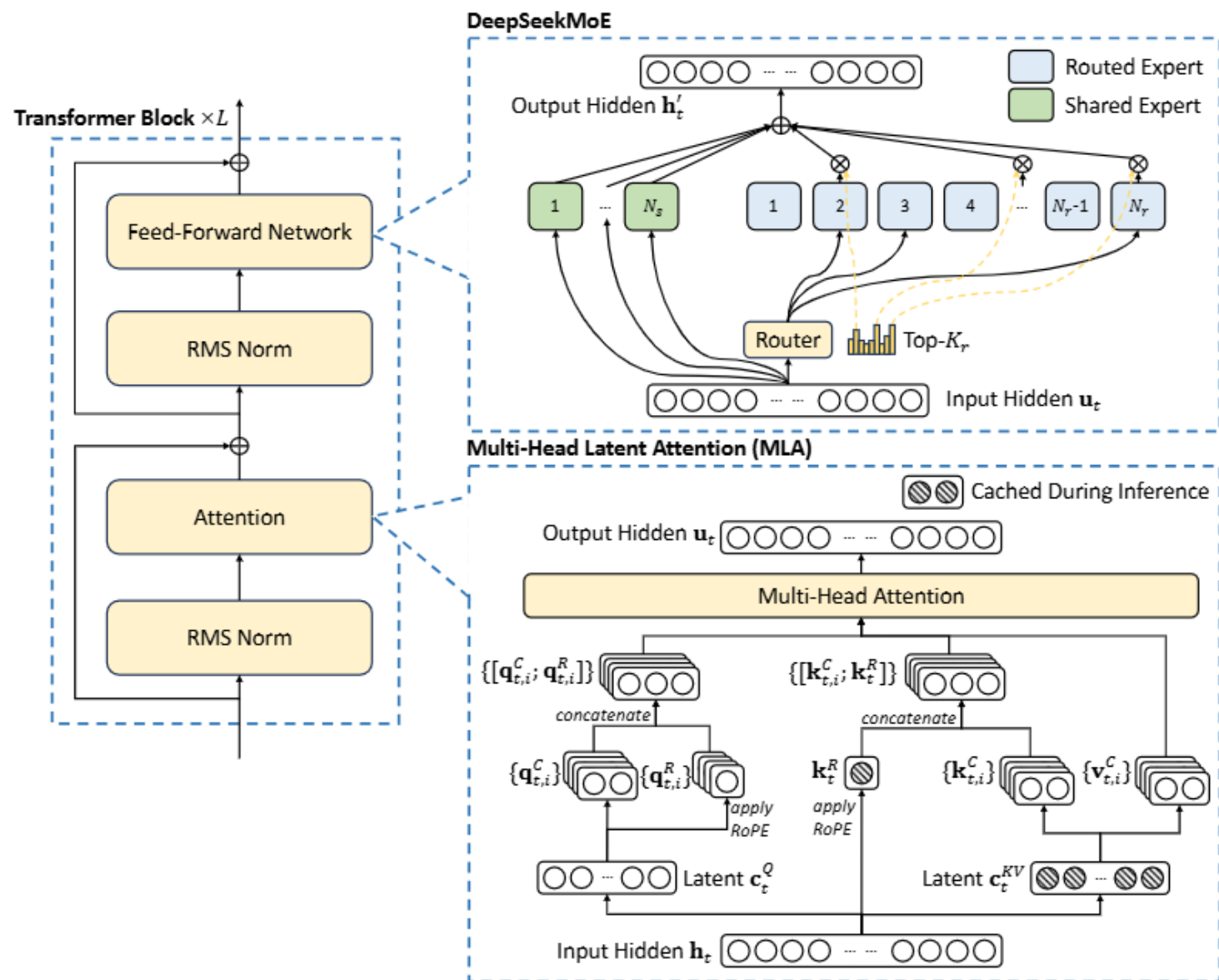




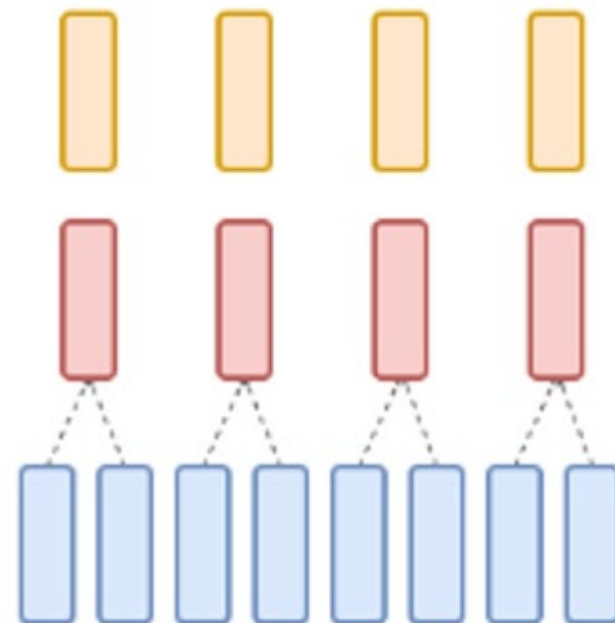
DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model



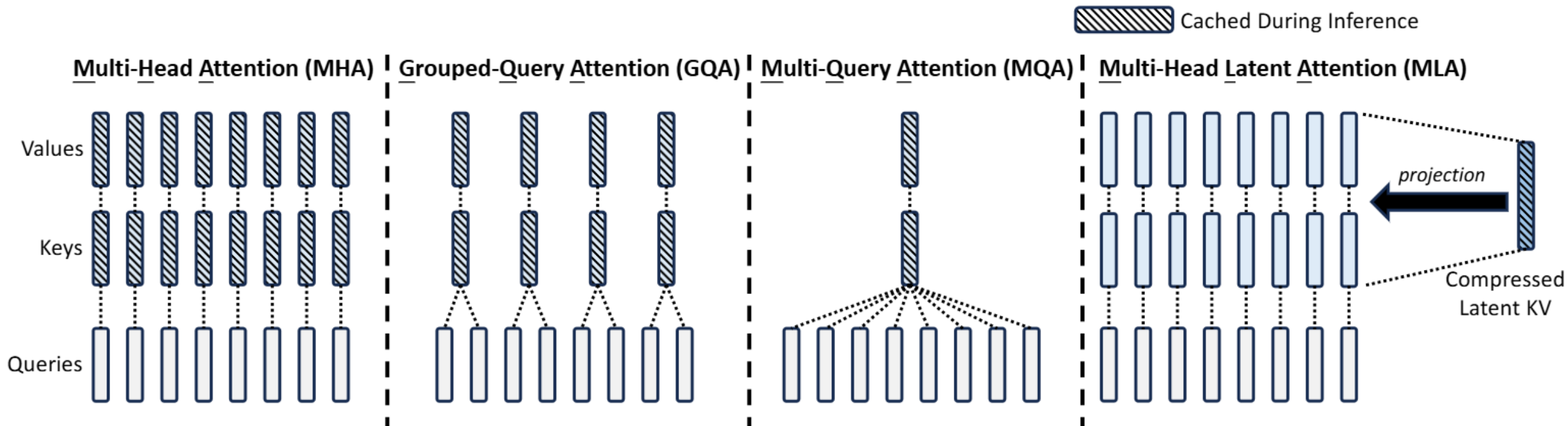
DeepSeek-V2



Grouped-query



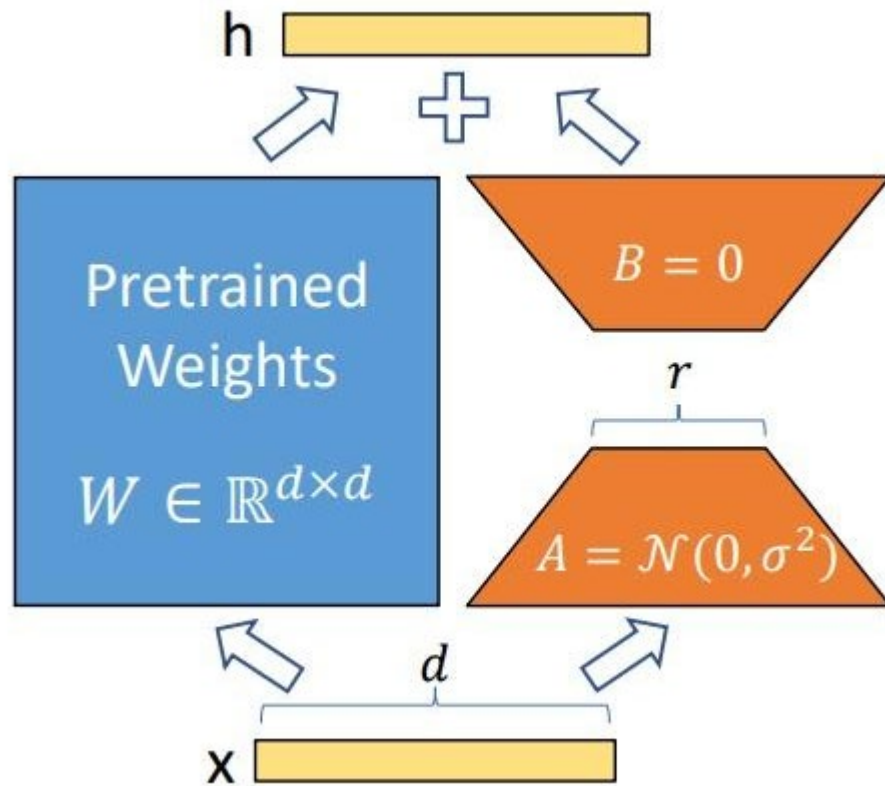
● Multi-Head Latent Attention (MLA)



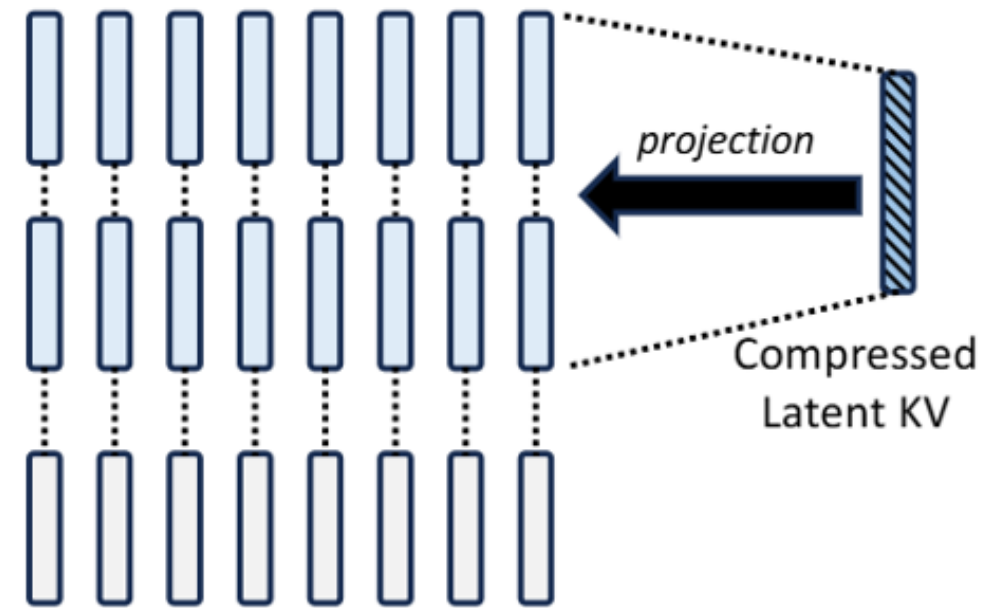
● 矩阵的秩

● Transformer的低秩性

● Multi-Head Latent Attention (MLA)

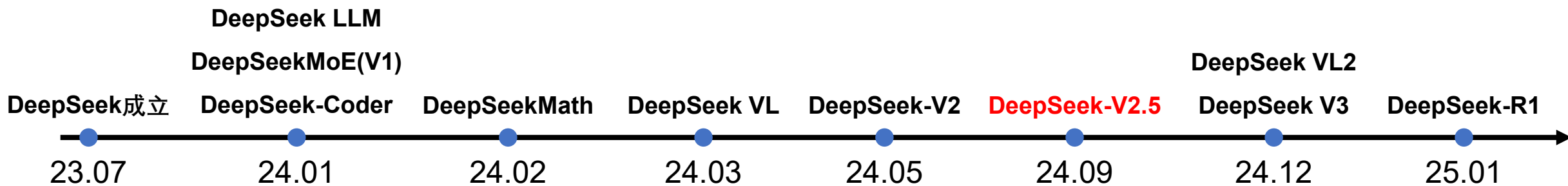


Multi-Head Latent Attention (MLA)



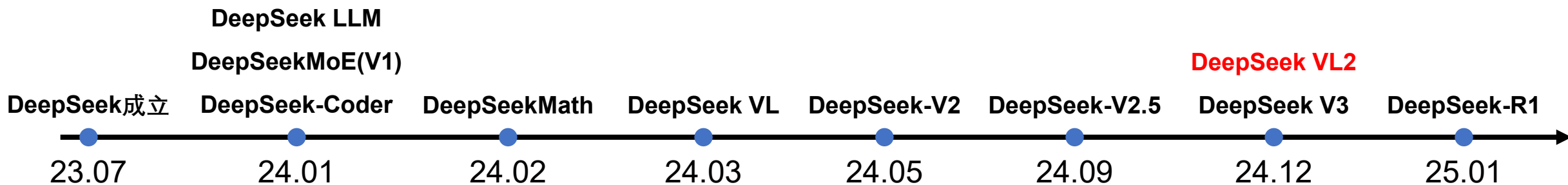


DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model



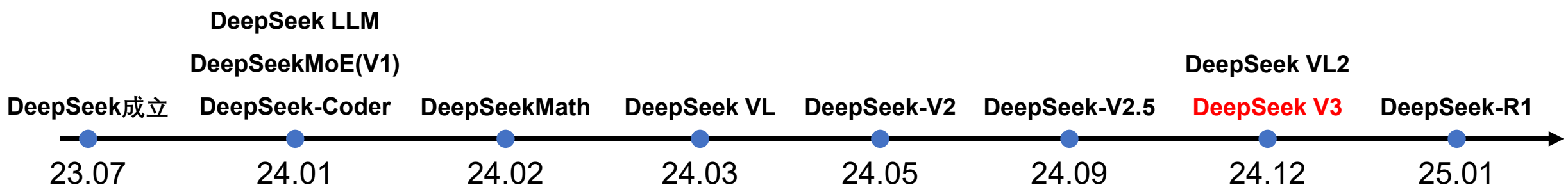


DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding

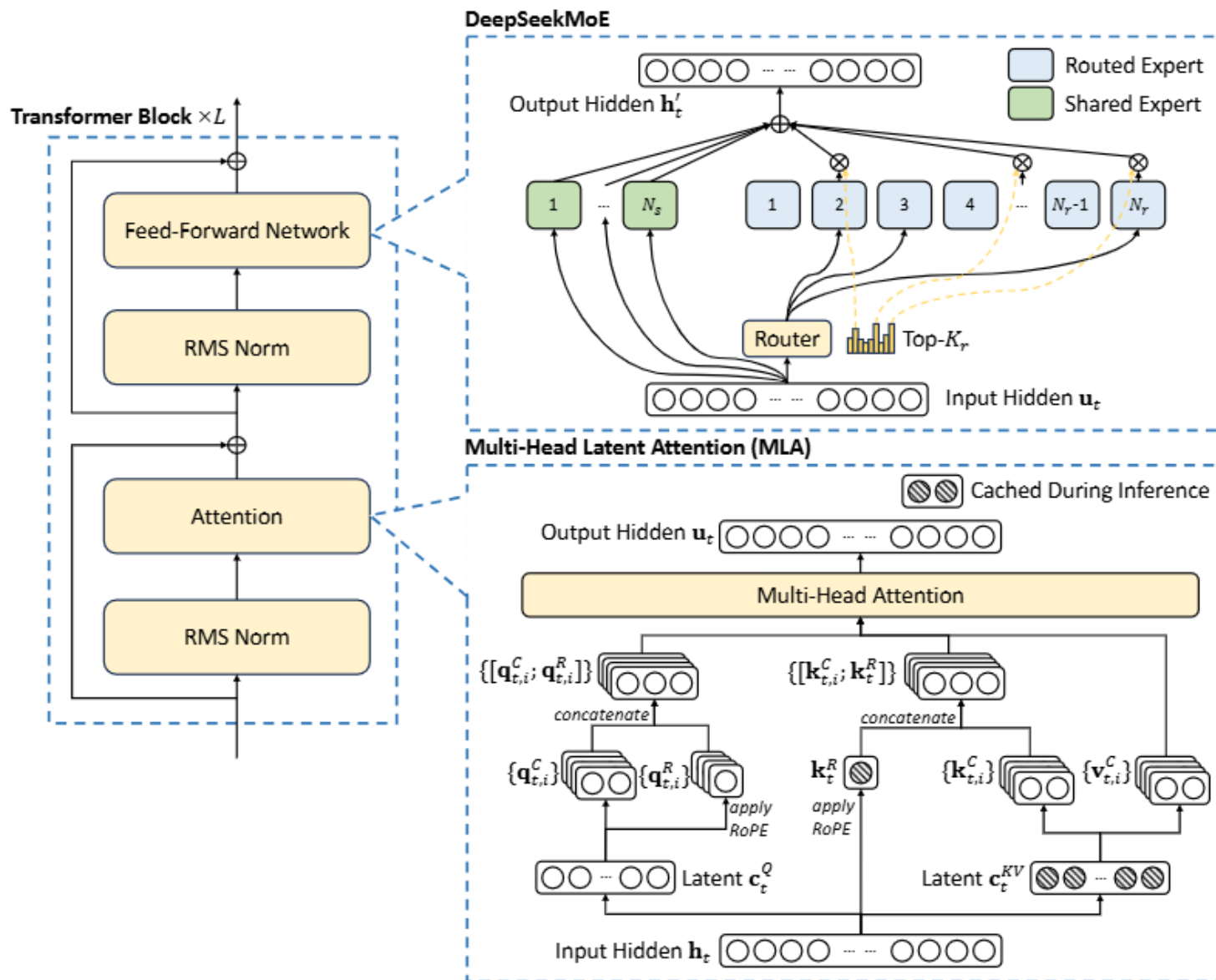




DeepSeek-V3 Technical Report



DeepSeek-V3



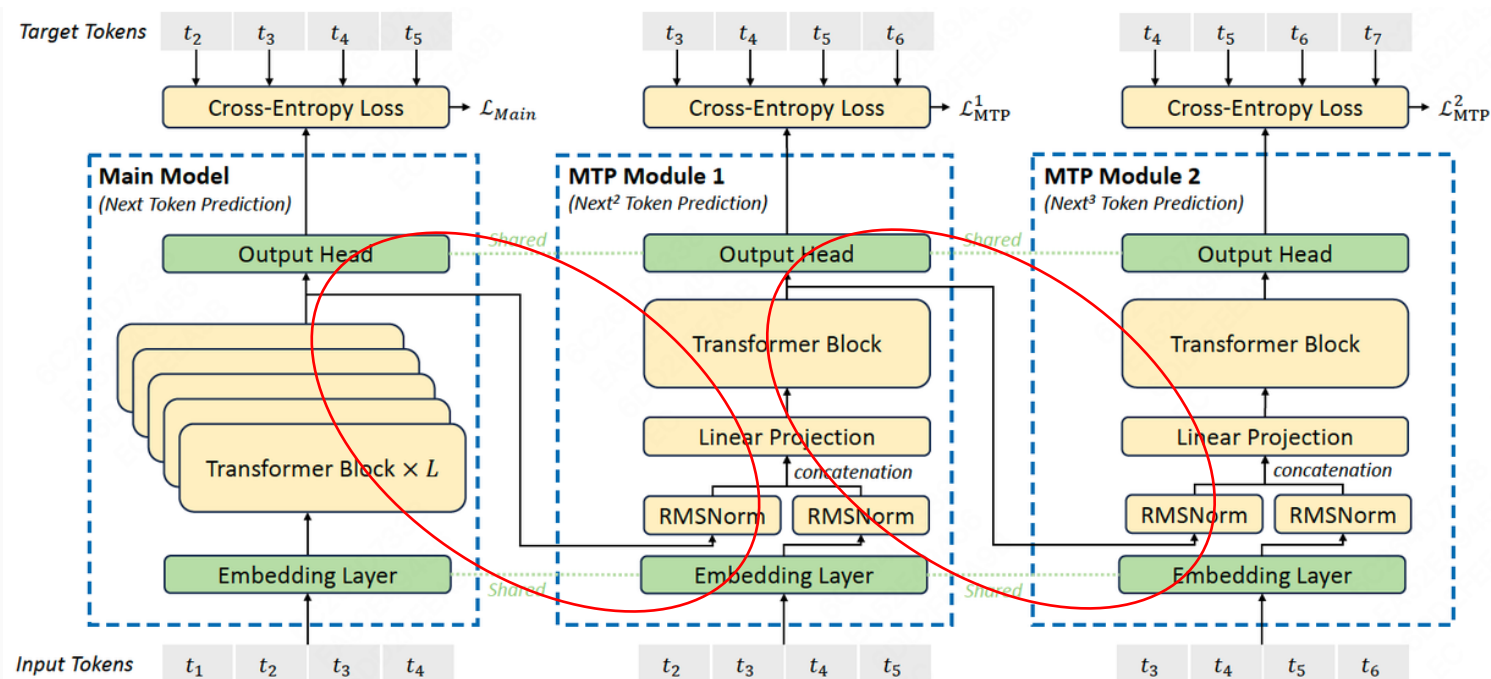
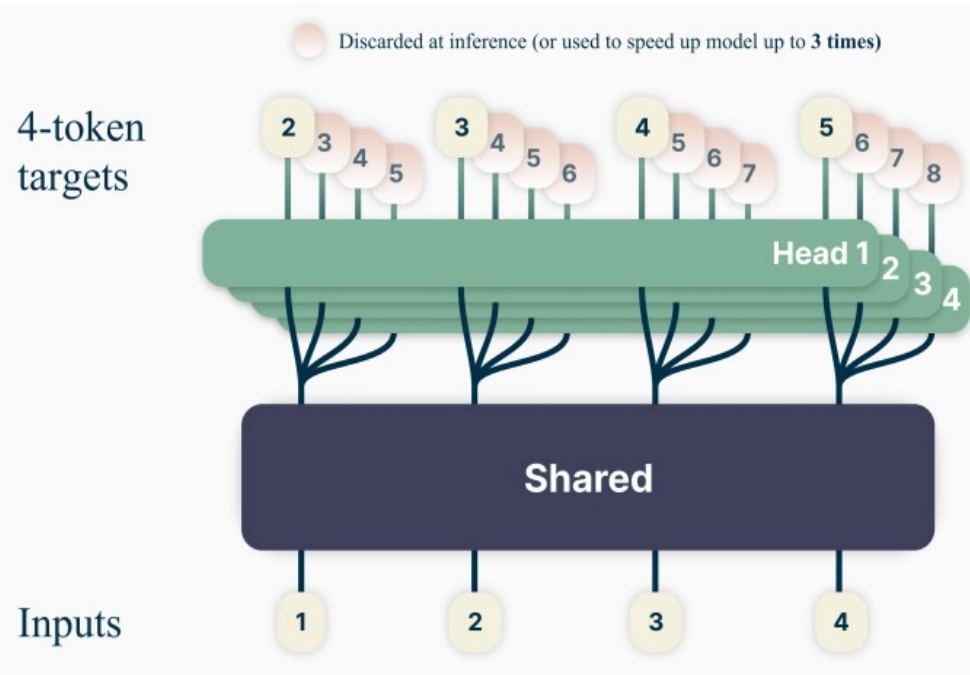
$$\mathbf{h}'_t = \mathbf{u}_t + \sum_{i=1}^{N_s} \text{FFN}_i^{(s)}(\mathbf{u}_t) + \sum_{i=1}^{N_r} g_{i,t} \text{FFN}_i^{(r)}(\mathbf{u}_t),$$

$$g_{i,t} = \frac{g'_{i,t}}{\sum_{j=1}^{N_r} g'_{j,t}},$$

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{Topk}(\{s_{j,t} | 1 \leq j \leq N_r\}, K_r), \\ 0, & \text{otherwise,} \end{cases}$$

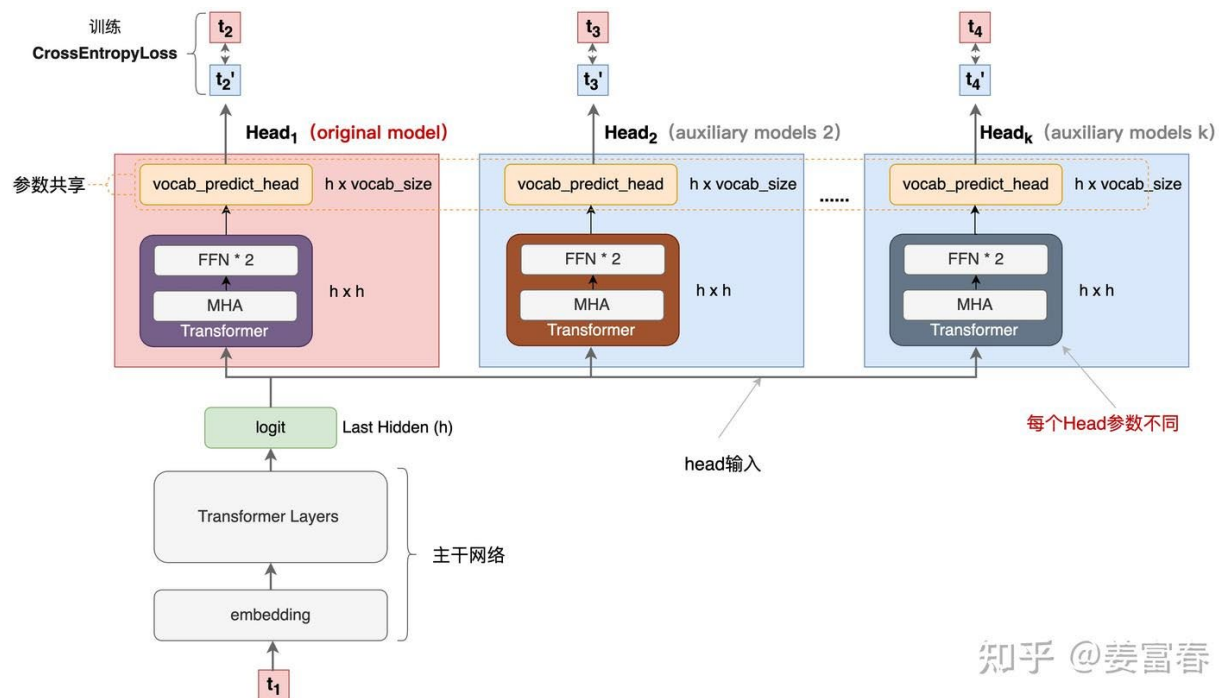
$$s_{i,t} = \text{Sigmoid}(\mathbf{u}_t^T \mathbf{e}_i),$$

● Multi-Token Prediction (MTP)

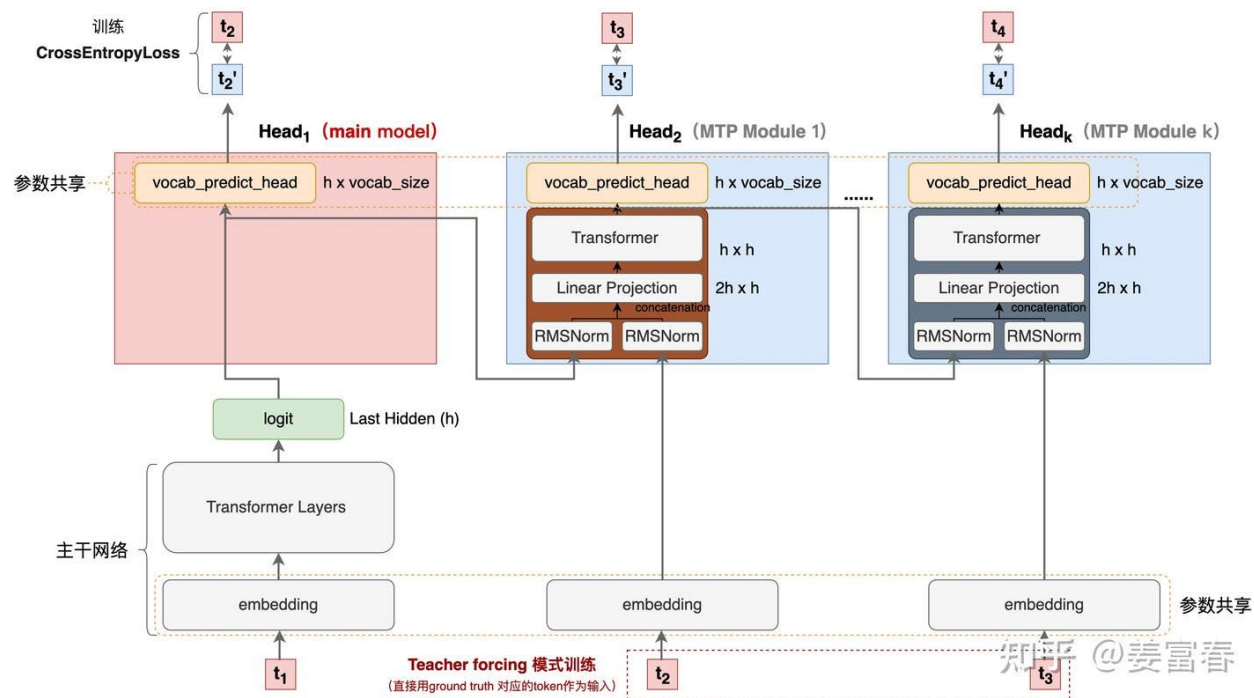


(causal chain)

Multi-Token Prediction (MTP)



Meta MTP



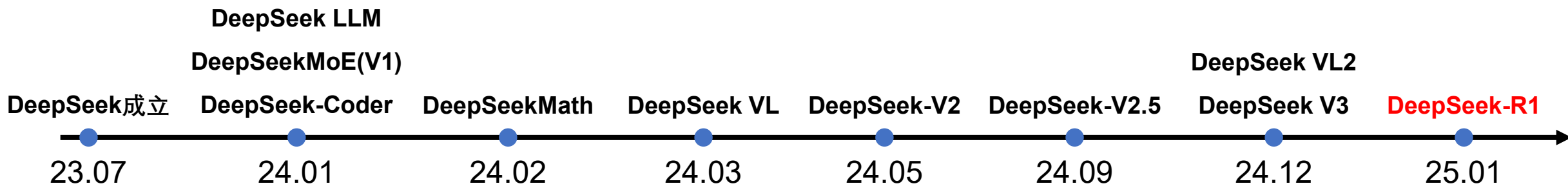
DeepSeek MTP

知乎 @姜富春

知乎 @姜富春



DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning



● 新一代推理大模型

在数学、逻辑推理、变成、科研问题问答等全方面推理性能全面对齐OpenAI o1模型，而且还能将其推理能力迁移至长文本创建、翻译、角色扮演领域。

目前全球综合能力（推理能力+文本问答能力）最好的开源模型

借助DeepSeek-R1蒸馏了出一组小尺寸模型并保留了其强大的推理能力，是目前业内最成功的蒸馏实验：例如1.5B的蒸馏模型编程能力超过了GPT4o

UC伯克利花4500美元的训练成本，制作DeepScaleR，1.5B的模型数学能力超越o1

● 三个贡献

1. 训练 DeepSeek-R1 Zero
2. 训练 DeepSeek-R1
3. 进行R1模型蒸馏

- **DeepSeek-R1 Zero: 证明了GRPO的有效性**

后训练 (post-training) : 再次训练已经训练好的模型以扩展用途

在DeepSeek-V3 Base的基础上获得

推理模型希望通过第二阶段的后训练, 让模型诞生思考的过程, 从而拥有推理能力

OpenAI 的o1通过延长“思维链 (Chain-of-Thought)”让模型拥有思维能力

另辟蹊径, 发现可以利用纯强化学习的方法GRPO让模型拥有推理能力

- **强化学习**

通过“试错”来让智能体学会如何在特定环境中做出最优决策

智能体、环境、状态、行动、奖励、新状态

基于价值的算法 (Value-based) 或基于策略的算法 (Policy-Based)

通常训练两个模型: 评价模型 (Critic model) 和策略模型 (Policy model)

● GRPO (Group Relative Policy Optimization)

两种奖励信号来源：准确性奖励 (Accuracy rewards) 和格式奖励 (Format rewards)

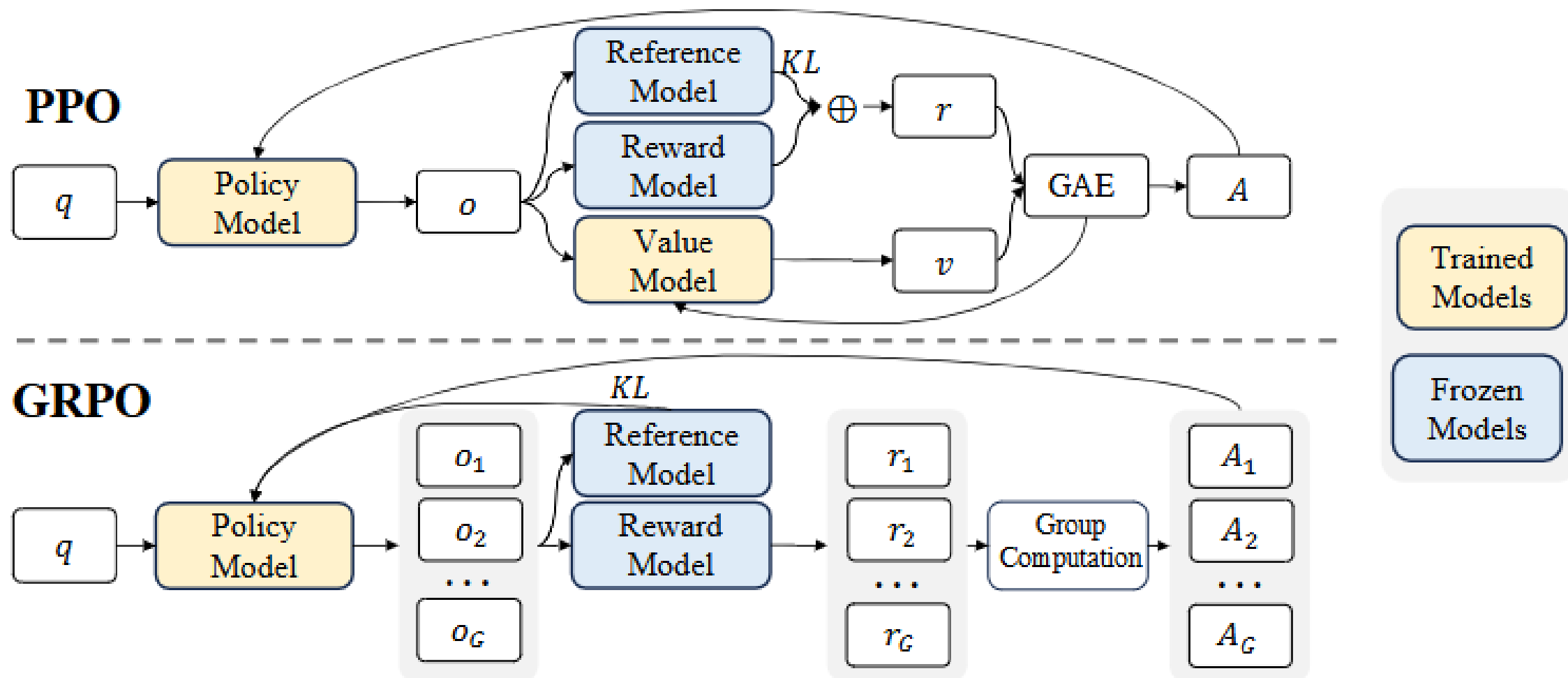
准确性奖励：用于评估响应是否正确，例如在数学问题中，要求模型能给出准确的最终答案。

格式奖励：要求答案和思考按照指定格式呈现，例如在框内。

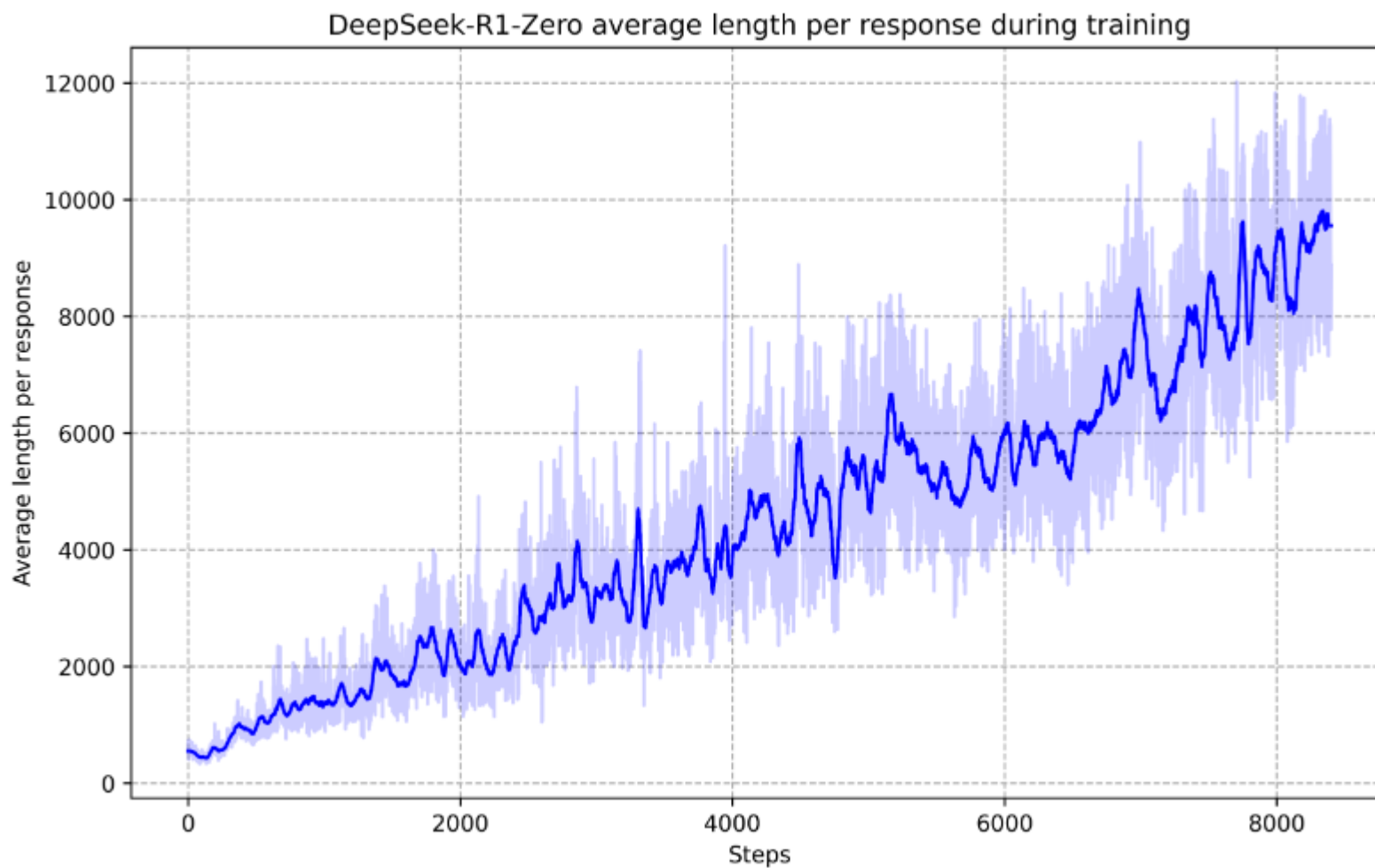
第一阶段：训练DeepSeek R1 Zero



- GRPO (Group Relative Policy Optimization)



- **DeepSeek-R1 Zero**



● 顿悟能力：aha时刻

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both \dots

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

\dots

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be \dots

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

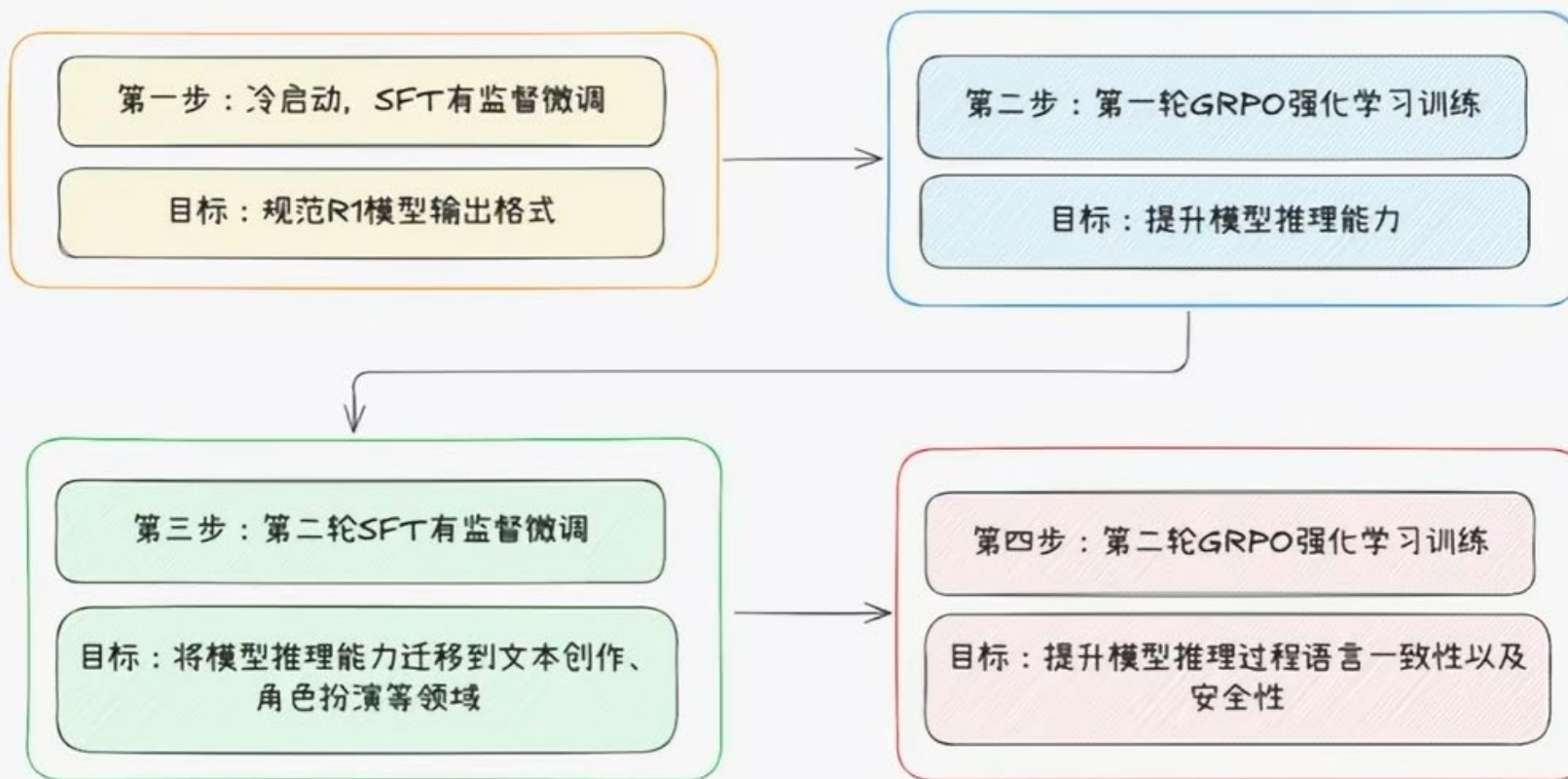
Next, I could square both sides again, treating the equation: \dots

\dots

● DeepSeek-R1

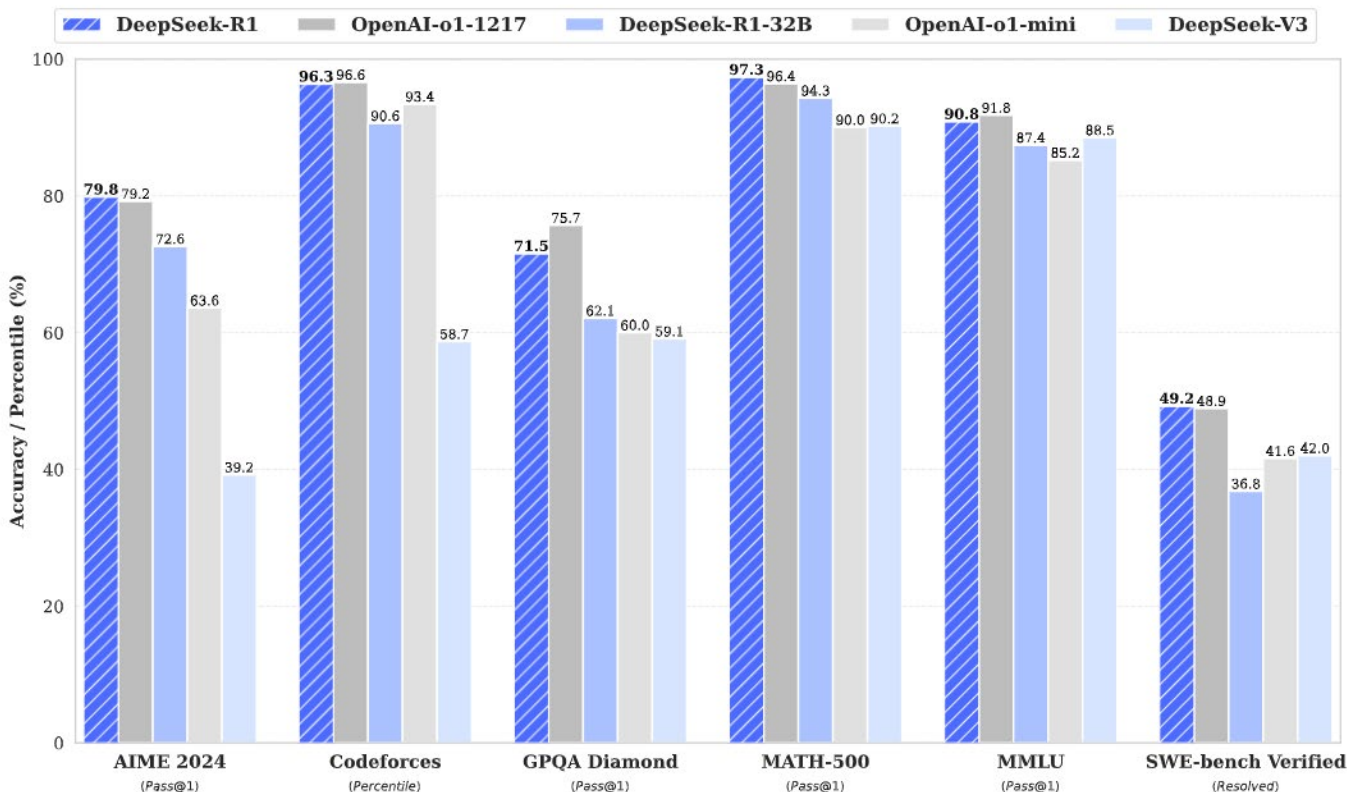
Deepseek R1 Zero缺陷：比如思维链可读性比差、语言混乱、模型不安全等

实验二：DeepSeek R1训练流程



● 蒸馏DeepSeek-R1

非常前沿的一个探索结论：纯强化学习所获得的推理能力是很容易通过蒸馏进行迁移的



实验一：训练R1 Zero

回顾&介绍GRPO
强化学习算法

训练DeepSeek R1 Zero模型

实验二：训练R1

结合R1-Zero训练
暴露的问题

训练DeepSeek R1模型

实验三：进行R1模型蒸馏

DeepSeek R1

DeepSeek R1 Distill Qwen
DeepSeek R1 Distill Llama



同濟大學
TONGJI UNIVERSITY

同心同德同舟楫
濟人濟事濟天下

感谢大家