2025

# Deepseek技术全景解析

PowerPoint Design

演讲人：楼立　　时间：2025.2

# 目录

CONTENTS


01 一、认识Deepseek

02 二、核心技术突破

03 三、私有化部署

04 四、性能优化参数

YOUR
LOGO

01

# 一、认识Deepseek

2025 Powerpoint Design

# Deepseek系列模型

## DeepSeekMoe模型

Deepseek- R1：670亿参数，推理优化，支持FP8量化，显存占用700G+。
Deepseek- V3：6710亿参数，MoE- 16架构，激活9专家。
Deepseek- Coder- V2：340亿参数，代码专项模型

## 多模态模型优势

- DeepSeek-VL2：4.5B,多模态版本，支持图像- 文本跨模态理解
- DeepSeek-VL2-Small: 2.8B, 支持图像- 文本跨模态理解
- DeepSeek-VL2-Tiny:1.0B, 支持图像- 文本跨模态理解

## R1 Distill系列模型

- DeepSeek-R1-Distill-Llama-70B
- DeepSeek-R1-Distill-Qwen-32B
- DeepSeek-R1-Distill-Qwen-14B
- DeepSeek-R1-Distill-Llama-8B
- DeepSeek-R1-Distill-Qwen-7B
- DeepSeek-R1-Distill-Qwen-1.5B

# 如何使用Deepseek模型能力

官方客户端/官网
https://chat.deepseek.com

## 官方开放平台接口调用

**1）API KEY申请：**

https://platform.deepseek.com/api_keys （新用户赠送50万token）

**2）官方API调用接口：**

```
import requests
headers = {"Authorization": "Bearer {API_KEY}"}
payload = {
    "model": "deepseek-r1",
    "messages": [{"role": "user", "content": "解释量子纠缠"}],
    "temperature": 0.7,
    "max_tokens": 512
}
response = requests.post("https://api.deepseek.com/chat/completions", json=payload, headers=headers)
```

**3）官方API文档：**

https://api-docs.deepseek.com/zh-cn/ （包含流式响应、函数调用等高级功能）

# 三方云平台访问Deepseek接口

**公有云API服务：**

**1）阿里百炼平台**

- **技术特性**：自研PPU芯片实现每token成本降低70%
- 模型版本：

```
- deepseek-r1:671B（128k输入/64k输出，QPS限制100）
- deepseek-v3:671B（支持多文档分析，最大并发32路）
```

**2）移动云-智算平台**

- **国产化方案**：昇腾910B+MindSpore架构，纯国产技术栈通过信创认证
- **战略合作**：提供定制化模型压缩服务（参数裁剪率可达50%）

**3）微软Azure云**

- **部署架构**：NVIDIA H200 + NVLink 4.0 + 200Gb InfiniBand
- **调用示例**：

```
response = client.chat.completions.create(
    model="deepseek-r1-8xH200",
    messages=[{"role": "user", "content": "生成Python数据可视化代码"}],
    temperature=0.5,
    stream=True  # 支持流式输出
)
```

YOUR
LOGO

02

# 二、核心技术突破

202X Powerpoint Design

# Deepseek的技术突破

## 无辅助损失负载均衡策略

1个共享专家，256个路由专家，每个token激活8-16个专家

动态专家路由系统通过计算专家权重，动态选择TopN专家，加权聚合输出。

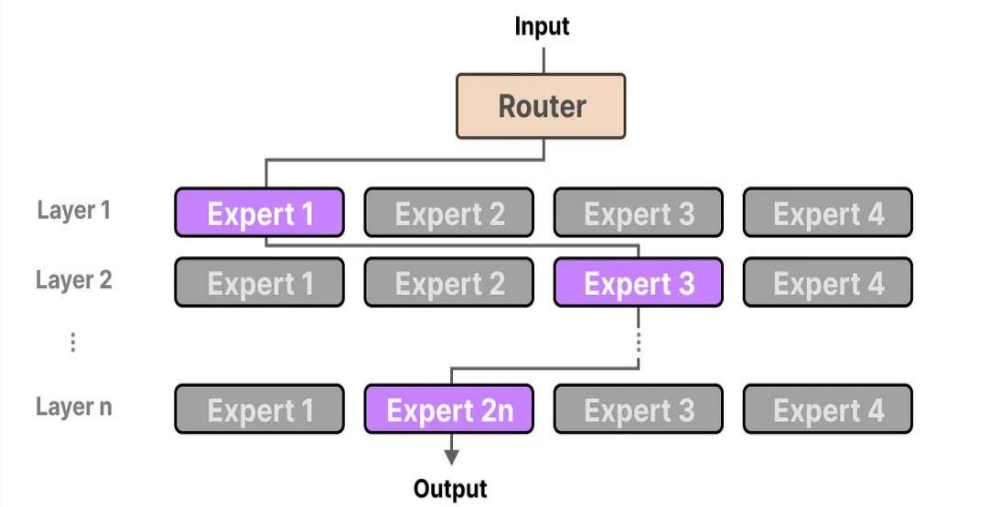
## 多tokens预测提升性能

置多tokens预测（MTP）训练目标可增强模型对未来多个tokens的预测能力，从而提升模型在各类任务上的整体性能，并能应用于推测解码以加速推理过程，提高生成效率。

## 架构与训练方法协同有效

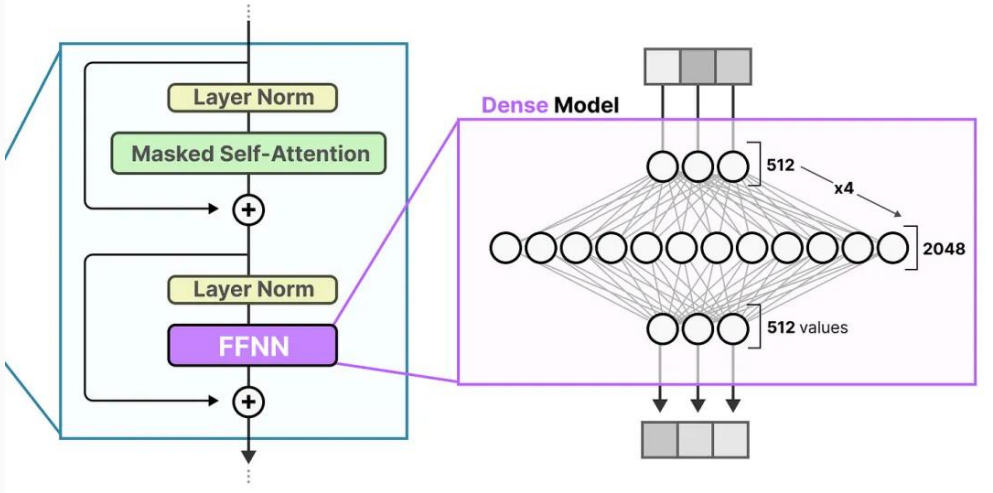
推测多头潜在注意力（MLA）、DeepSeekMoE 架构以及 FP8 混合精度训练等方法相互兼容且协同作用，能够在大规模模型训练中实现高效训练和出色性能表现，共同推动模型能力提升。

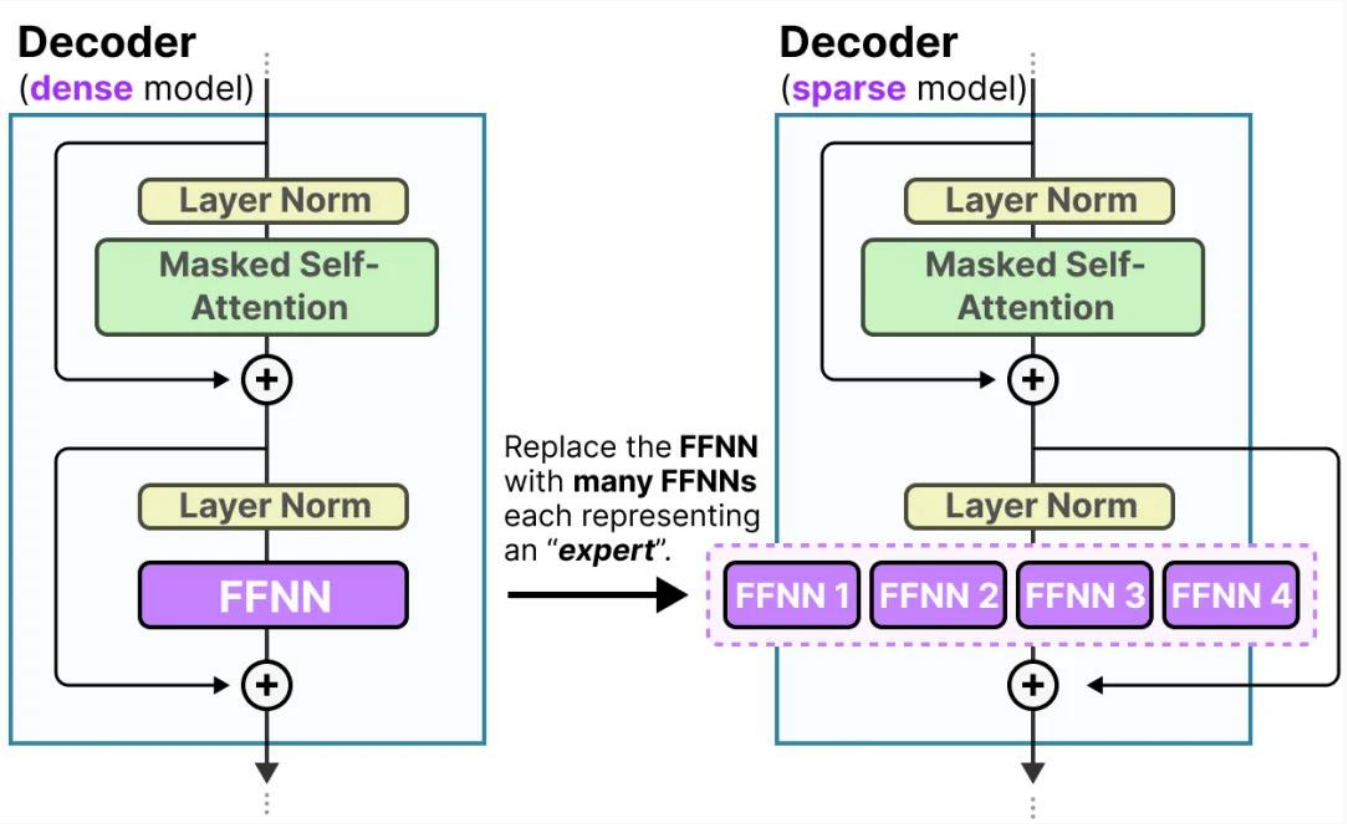
# DeepseekMoE（稀疏） VS Dence（密集）




密集模型：全连接，计算量大

稀疏模型：分区域，选择部分区域，区域内全连接，大幅优化计算量



# 无辅助损失负载均衡策略

## DeepSeekMoE 与负载均衡

- **基本架构**：在 Feed - Forward Networks (FFNs)中采用 DeepSeekMoE 架构，使用更精细的专家设置，计算 FFN 输出$h_t'$的公式为$h_t' = u_t + \sum_{i=1}^{N_s} FFN_i^{(s)}(u_t) + \sum_{i=1}^{N_r} g_{i,t} FFN_i^{(r)}(u_t)$，其中$N_s$和$N_r$分别是共享专家和路由专家的数量，$g_{i,t} = \frac{g'_{i,t}}{\sum_{j=1}^{N_r} g'_{j,t}}$，
$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in Topk(\{s_{j,t} | 1 \le j \le N_r\}, K_r) \\ 0, & otherwise \end{cases}$，$s_{i,t} = Sigmoid\left(u_t^T e_i\right)$。
- **无辅助损失负载均衡**：引入偏差项$b_i$到亲和分数$s_{i,t}$来确定 top - K 路由，即
$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} + b_i \in Topk(\{s_{j,t} + b_j | 1 \le j \le N_r\}, K_r) \\ 0, & otherwise \end{cases}$，在训练过程中动态调整偏差项以保持专家负载平衡。同时采用互补的sequence-wise辅助损失$\mathcal{L}_{Bal} = \alpha \sum_{i=1}^{N_r} f_i P_i$，其中$f_i = \frac{N_r}{K_r T} \sum_{t=1}^{T} \mathbf{1}(s_{i,t} \in Topk(\{s_{j,t} | 1 \le j \le N_r\}, K_r))$，$s'_{i,t} = \frac{s_{i,t}}{\sum_{j=1}^{N_r} s_{j,t}}$，$P_i = \frac{1}{T} \sum_{t=1}^{T} s'_{i,t'}$，$\alpha$为平衡因子。
- **节点限制路由**：采用限制路由机制，确保每个tokens最多发送到$M$个节点，根据节点上专家的亲和分数总和选择节点，以限制通信成本，实现计算 - 通信重叠。

# FP8混合精度训练

**FP8 训练方法**

- **细粒度量化**：对激活和权重采用不同的分组量化方式，激活按$1x128$ tile 分组，权重按$128x128$块分组，并引入组缩放因子，以更好地处理特征异常值，提高量化精度。
- **提高积累精度**：针对 FP8 GEMM 运算在 NVIDIA H800 GPUs 上积累精度有限的问题，采用在 Tensor Cores 上执行一定间隔后将中间结果复制到 CUDA Cores 进行全精度 FP32 积累的策略，如设置$N_C = 128$元素可显著提高精度且不引入过多开销。
- **其他策略**：采用 E4M3 格式在所有张量上提高精度，通过在线计算缩放因子进行量化，并对不同算子的存储和通信格式进行优化，如采用 BF16 跟踪 AdamW 优化器的部分参数，对某些算子的激活采用定制格式等。
- **混合精度框架**：多数计算密集型操作采用 FP8 精度，如 GEMM 操作，同时对部分关键操作（如嵌入模块、输出头、MoE 门控模块、归一化和注意力算子等）保留原精度（如 BF16 或 FP32），以平衡训练效率和数值稳定性。
- **精度提升策略**

YOUR
LOGO

03

# 三、私有化部署

# Deepseek满血版

202X Powerpoint Design

# 模型文件下载

HuggingFace：
https://huggingface.co/deepseek-ai/DeepSeek-R1/tree/main

| | | |
|---|---|---|
| LICENSE Safe | 1.06 kB | |
| README.md Safe | 19 kB | |
| config.json Safe | 1.73 kB | |
| configuration_deepseek.py Safe | 10.6 kB | |
| generation_config.json Safe | 171 Bytes | |
| model-00001-of-000163.safetensors Safe | 5.23 GB LFS | |
| model-00002-of-000163.safetensors Safe | 4.3 GB LFS | |
| model-00003-of-000163.safetensors Safe | 4.3 GB LFS | |
| model-00004-of-000163.safetensors Safe | 4.3 GB LFS | |
| model-00005-of-000163.safetensors Safe | 4.3 GB LFS | |
| model-00006-of-000163.safetensors Safe | 4.37 GB LFS | |
| model-00007-of-000163.safetensors Safe | 4.31 GB LFS | |

魔塔社区：
https://modelscope.cn/models/unsloth/DeepSeek-R1-BF16/files

| | |
|---|---|
| .gitattributes | 1.52KB |
| config.json | 1.61KB |
| configuration.json | 64.00B |
| configuration_deepseek.py | 10.56KB |
| generation_config.json | 171.00B |
| LICENSE | 1.06KB |
| model-00001-of-000163.safetensors Safetensors | 8.61GB |
| model-00002-of-000163.safetensors Safetensors | 8.60GB |
| model-00003-of-000163.safetensors Safetensors | 8.60GB |

# 硬件选型建议

## GPU选型与数据格式

**FP8适配方案：**

| 供应商 | 硬件配置说明 |
| --- | --- |
| 英伟达、微软AZure | NVIDIA H200集群：单卡141G，最小部署配置8卡，单台，建议NVLINK |
| 火山引擎、UCloud | NVIDIA H20集群：单卡96G，最小部署配置8卡，两台，建议NVLINK+InfinitBand |
| | AMD MI300X：单卡支持160GB显存，8卡可部署 |

**BF16适配方案：**

| 供应商 | 硬件配置说明 |
| --- | --- |
| | NVIDIA A100机器：A100，A800，4台8卡可部署，建议NVLINK+InfinitBand |
| 移动云、华为云 | 华为昇腾910B：单卡64G，4台8卡可部署 |
| 阿里云 | 阿里PPU 810E：单卡96G，1台16卡可部署 |
| Groq | LPU：Language Processing Unit，推理专用芯片，Groq LPU |

# 推理框架配置

## SGLang（推荐）

SGLang 目前支持 MLA 优化、DP 注意力机制、FP8（W8A8）、FP8 KV 缓存和 Torch 编译，在开源框架中实现了最先进的延迟和吞吐量性能。

值得注意的是，SGLang v0.4.1 完全支持在 NVIDIA 和 AMD 的 GPU 上运行 DeepSeek-V3，使其成为一种高度通用且强大的解决方案。

SGLang 还支持多节点张量并行，使您能够在多台联网的机器上运行此模型。

多标记预测（MTP）正在开发中，其进展情况可在优化计划中进行追踪。

## vLLM

vLLM是一个业界流行的用于LLM推理加速的库

Support DeepSeek-V3 model with FP8 and BF16 modes for tensor parallelism and pipeline parallelism.

## ollama

仅实用本机部署

不适合作为生产推理框架

YOUR
LOGO

04

# 四、性能调优

202X Powerpoint Design

# 黄金参数组合

## 事实性问答参数

Temperature：0.3，Top_p：0.9，Frequency Penalty：0.7，准确率提升12%。

优化参数组合，提升模型在事实性问答场景中的准确率。

事实性问答场景对模型准确性要求高，优化参数组合效果显著。

## 创意生成参数

Temperature：0.8，Top_p：0.7，Frequency Penalty：0.3，多样性指数增加45%。

优化参数组合，提升模型在创意生成场景中的多样性。

创意生成场景对模型多样性要求高，优化参数组合效果显著。

## 代码生成参数

Temperature：0.5，Top_p：0.85，Frequency Penalty：0.5，HumanEval通过率+18%。

优化参数组合，提升模型在代码生成场景中的通过率。

代码生成场景对模型生成质量要求高，优化参数组合效果显著。

# 推理训练参数优化

| 参数类别 | 关键参数 | 说明 |
| --- | --- | --- |
| 专家模型 | `num_experts=8` | 激活专家数量，影响计算资源和推理质量平衡 |
| | `expert_dropout=0.1` | 专家随机丢弃率，防止过拟合 |
| 并行策略 | `tensor_parallel=4` | 张量并行度，需与GPU数量匹配 |
| | `pipeline_parallel=2` | 流水线并行级数，适合长序列处理 |
| 窗口优化 | `context_window=131072` | 最大上下文token数，超过将触发滚动缓存机制 |
| | `window_overlap=512` | 上下文滚动重叠区域，保持语义连贯性 |
| 计算优化 | `fp8_quantization=True` | 启用FP8量化推理，显存占用减少40% |
| | `kv_cache_ratio=0.3` | KV缓存压缩率，平衡内存与计算效率 |

DESIGN

2025

# 感谢Deepseek 让AI离我们更近

PowerPoint Design

演讲人：楼立　　时间：2025.2