EnQ Digital

INTELIGÊNCIA ARTIFICIAL | ARQUITETURA LLM

Players LLM x MoE: quem já está usando Mixture of Experts?

Mapa executivo dos modelos que adotam arquitetura de especialistas, com leitura técnica e implicações para infraestrutura, custo e estratégia.

Resumo executivo

Por que MoE esta virando uma arquitetura central em LLMs

A arquitetura **Mixture of Experts** (MoE) deixou de ser apenas uma técnica academica e passou a aparecer em modelos comerciais, open-weight e enterprise. O motivo e simples: ela permite aumentar a capacidade total do modelo sem ativar todos os parâmetros em cada token. Na pratica, isso abre caminho para modelos maiores, com melhor relação entre qualidade, custo de inferência e throughput.

Este e-book mapeia os principais players que publicamente associaram seus modelos a MoE, explica como a arquitetura funciona e traduz o tema para estratégia de infraestrutura, LLMaaS, agentes, data center e operação empresarial.

IDEIA-CHAVE

Sparse compute

Experts ativos por token

EFEITO PRATICO

Mais capacidade

Total maior, ativo menor

DECISAO TÉCNICA

Infra importa

Custo real depende da stack

Leitura executiva
MoE nao significa automaticamente “melhor modelo”. Significa uma forma diferente de escalar: parte do modelo fica especializada e apenas uma fração e acionada por token. O ganho depende de treinamento, roteamento, dados, balanceamento dos especialistas, memória, rede e stack de serving.

Status dos players: Google, Meta, Mistral, Alibaba/Qwen, DeepSeek, xAI, Databricks, Snowflake, AI21, Moonshot/Kimi e MiniMax possuem divulgações públicas associando modelos relevantes a MoE. OpenAI e Anthropic sao tratados neste material como **N/D publicamente**, pois nao confirmam abertamente a arquitetura de seus modelos frontier atuais.

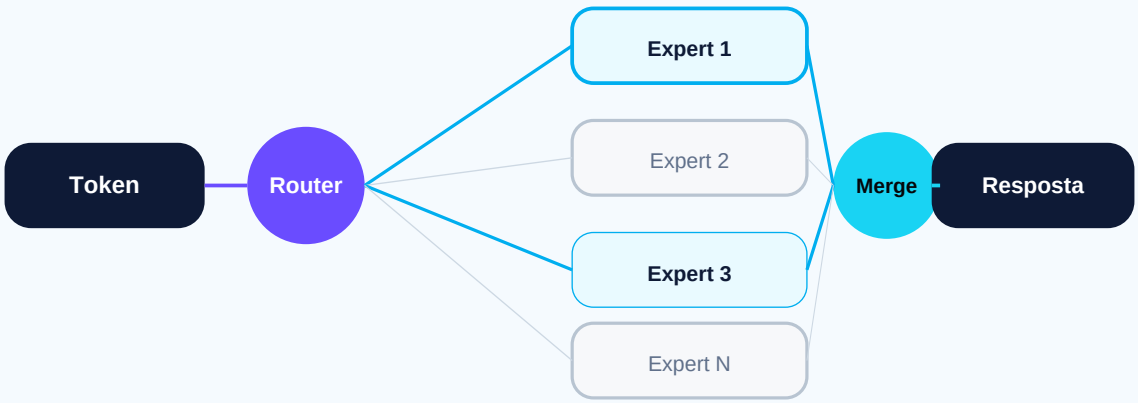
Sumario

Mapa rapido do e-book

1. O que e MoE em LLMs

A arquitetura de especialistas explicada sem simplificar demais

Em um modelo denso tradicional, a maior parte dos pesos relevantes participa do processamento de cada token. Em um modelo MoE, certas camadas sao substituidas por um conjunto de **especialistas**, normalmente redes feed-forward especializadas. Um **router** escolhe quais especialistas devem processar cada token, frequentemente usando top-1, top-2, top-4 ou top-k.



O roteador escolhe poucos especialistas por token. O custo acompanha os parâmetros ativos, nao o total.

A diferença conceitual mais importante e separar **parâmetros totais** de **parâmetros ativos**. Um modelo pode ter centenas de bilhões ou ate trilhões de parâmetros totais, mas ativar apenas uma fração por token. Isso reduz o custo computacional por passo de inferência, embora a memória, a rede e a orquestração ainda precisem lidar com o modelo como um sistema distribuido.

Formula mental
Dense escala ativando quase tudo. MoE escala aumentando a capacidade total, mas acionando especialistas conforme o token e a tarefa. Em vez de “um cerebro gigante sempre inteiro”, pense em “uma banca de especialistas com roteador”.

Componentes centrais:

- **Router/Gate:** calcula quais especialistas devem receber cada token.
- **Experts:** sub-redes especializadas, geralmente MLPs/feed-forward, com pesos independentes.
- **Top-k:** numero de especialistas selecionados por token.
- **Load balancing:** mecanismo para evitar que poucos experts fiquem sobrecarregados.
- **Expert parallelism:** técnica para distribuir especialistas entre GPUs/nos.

2. Linha do tempo: de pesquisa a produção

MoE amadureceu porque a pressao por eficiência virou decisiva

A ideia de computação condicional antecede os LLMs modernos, mas ganhou tração com a camada *Sparsely-Gated Mixture-of-Experts* proposta em 2017, que mostrou como aumentar capacidade sem crescimento proporcional de custo computacional. Depois, trabalhos como Switch Transformer simplificaram o roteamento e demonstraram modelos esparsos em escala de trilhões de parâmetros.

Periodo	Marco	Impacto
2017	Sparsely-Gated MoE	Populariza a camada de especialistas com gate treinavel e computação condicional.
2021-2022	Switch Transformer / GShard	Simplificação do roteamento e prova de escala em modelos muito grandes.
2023-2024	Mixtral, Grok-1, DBRX, Arctic, Jamba	MoE passa a aparecer em modelos abertos, enterprise e APIs comerciais.
2025-2026	DeepSeek-V3, Qwen3, Llama 4, Kimi K2, MiniMax	A disputa passa a envolver parâmetros ativos, custo de inferência, agentes e long context.

O ponto critico para empresas e que MoE nao e apenas uma escolha de arquitetura de modelo. Ele tambem altera o desenho da infraestrutura: mais dependencias entre GPUs, maior importancia de rede de baixa latência, necessidade de orquestrar especialistas, e maior complexidade no monitoramento de latência por token.

3. Players LLM x MoE

Quem aparece publicamente associado a Mixture of Experts

A tabela abaixo separa modelos com MoE publicamente declarado daqueles cujo uso de MoE nao foi confirmado. Para modelos fechados, o fato de haver rumores no mercado nao deve ser tratado como evidencia técnica. Em posicionamento corporativo, o correto e usar “confirmado publicamente” ou “nao divulgado”.

Player	Modelo/familia	MoE publico	Parametros divulgados	Leitura estratégica
Google	Gemini 1.5	Sim	Arquitetura MoE anunciada; detalhes de parâmetros nao abertos.	MoE em modelo comercial de grande escala e contexto longo.
Meta	Llama 4 Scout / Maverick	Sim	Scout: 17B ativos, 16 experts; Maverick: 17B ativos, 128 experts.	Open-weight multimodal com MoE, forte para ecossistema e integradores.
Mistral AI	Mixtral 8x7B / 8x22B	Sim	8x7B: 47B totais, 13B ativos.	Referencia open-weight para uso eficiente e deployment flexivel.
Alibaba / Qwen	Qwen3-235B-A22B / 30B-A3B	Sim	235B totais, 22B ativos; 30B totais, 3B ativos.	MoE aberto com foco em raciocínio, código e multilingue.
DeepSeek	DeepSeek-V3	Sim	671B totais, 37B ativos por token.	Exemplo de escala agressiva com custo ativo menor.
xAI	Grok-1	Sim	314B parâmetros; 25% ativos por token.	Open-weight grande, relevante como referência de MoE fechado/aberto.
Databricks	DBRX	Sim	132B totais, 36B ativos; 16 experts, escolhe 4.	MoE enterprise integrado a stack de dados e treinamento corporativo.
Snowflake	Arctic	Sim	480B totais, 17B ativos; Dense-MoE hybrid.	LLM enterprise com foco em eficiência e workloads de dados.
AI21	Jamba / Jamba 1.5	Sim	Arquitetura hibrida Transformer-Mamba-MoE.	Long context e eficiência com design hibrido.
Moonshot AI	Kimi K2	Sim	1T total, 32B ativos.	MoE voltado a agentes, código e tool-use em escala aberta.
MiniMax	MiniMax-Text-01 / M1	Sim	456B totais, 45.9B ativos.	Long context e raciocínio com arquitetura hibrida + MoE.
OpenAI	GPT-4o / GPT-4.1 / outros	N/D	Arquitetura nao divulgada publicamente.	Evitar afirmar MoE sem confirmação oficial.
Anthropic	Claude	N/D	Arquitetura nao divulgada publicamente.	Tratar como modelo fechado sem especificação publica de MoE.

Observação: os numeros acima devem ser usados como fotografia de mercado baseada em divulgações públicas. Em modelos fechados, rumores de arquitetura nao substituem fonte oficial, paper técnico, model card ou repositorio publicado pelo fornecedor.

4. Leitura dos principais players

O que cada movimento indica para o mercado

Google - Gemini 1.5

O Google afirmou que Gemini 1.5 usa uma nova arquitetura MoE para melhorar eficiência de treinamento e serving. O caso e estratégico porque associa MoE a modelos multimodais e de contexto longo em ambiente comercial via Google AI Studio e Vertex AI.

Meta - Llama 4

A Meta declarou que Llama 4 Scout e Maverick sao seus primeiros modelos Llama construidos com MoE. A combinação de multimodalidade, open-weight e especialistas coloca a familia Llama em uma posição importante para integradores.

Mistral - Mixtral

Mixtral 8x7B popularizou MoE open-weight. O modelo tem 47B parâmetros totais e 13B ativos, tornando-se um exemplo didatico de como MoE separa capacidade total e custo ativo.

DeepSeek - V3

DeepSeek-V3 e um dos exemplos mais citados de escala MoE: 671B parâmetros totais e 37B ativados por token. O interesse do mercado vem da relação entre performance, custo e arquitetura aberta o suficiente para analise técnica.

Qwen - Qwen3 MoE

A familia Qwen3 trouxe variantes densas e MoE. O Qwen3-235B-A22B, com 235B totais e 22B ativos, tornou-se referência para uso open-weight em raciocínio, código e aplicações multilingues.

Databricks e Snowflake

DBRX e Arctic mostram que MoE tambem e tema enterprise: nao apenas para chat, mas para dados, SQL, RAG, governança, treinamento customizado e modelos corporativos.

Moonshot/Kimi e MiniMax

Kimi K2 e MiniMax-M1/Texto-01 mostram uma tendencia forte em modelos abertos e asiaticos: MoE, long context, foco em agentes, código e tool-use com parâmetros ativos controlados.

5. Dense x MoE

Comparação executiva para escolher arquitetura ou fornecedor

A decisão entre modelos densos e MoE nao e binaria. Em muitos ambientes, o melhor portfólio combina modelos densos pequenos para tarefas de baixa complexidade, modelos densos grandes para previsibilidade e modelos MoE para workloads de alta capacidade, agentes, código, long context ou cenarios de alto throughput.

Criterio	Modelo denso	Modelo MoE
Custo por token	Mais diretamente proporcional ao tamanho ativo do modelo.	Pode ser menor que o tamanho total sugeriria, pois ativa poucos especialistas.
Memoria	Normalmente precisa carregar o modelo completo, mas arquitetura e serving sao mais simples.	Tambem precisa gerenciar muitos pesos; memória e distribuição dos experts importam muito.
Latencia	Mais previsivel em lotes pequenos.	Pode oscilar por roteamento, balanceamento, comunicação all-to-all e cache.
Qualidade	Boa estabilidade e simplicidade operacional.	Pode ganhar capacidade e especialização, mas depende de roteamento e treinamento bem balanceados.
Operação	Mais facil para times menores.	Exige stack madura: expert parallelism, observabilidade e rede de alta performance.
Melhor uso	Classificação, resumo, atendimento, tarefas padronizadas, edge ou menor custo.	Agentes, código, reasoning, long context, throughput alto e modelos maiores.

Erro comum
Nao venda MoE como “IA que pensa com varios cerebros” sem explicar custo, roteamento e parâmetros ativos. A narrativa correta para negocios e eficiência de escala: mais capacidade potencial com computação seletiva.

6. Infraestrutura para rodar MoE

Onde o ganho aparece - e onde a complexidade aparece tambem

Do ponto de vista de infraestrutura, MoE e atraente porque reduz os parâmetros ativos por token, mas nao elimina a necessidade de memória para pesos, rede de alta velocidade e orquestração. O custo real depende de batch size, tamanho de contexto, precisão dos pesos, quantização, KV cache, comunicação entre GPUs e distribuição dos experts.

Camadas de infraestrutura critica:

- **GPU e memória:** modelos com centenas de bilhões de parâmetros exigem memória agregada, mesmo quando poucos parâmetros sao ativos por token.
- **Rede de baixa latência:** expert parallelism pode gerar comunicação all-to-all entre GPUs e nos.
- **Serving engine:** vLLM, TensorRT-LLM, SGLang, TGI e stacks proprietarias precisam suportar roteamento eficiente.
- **Quantização:** FP8, INT8, INT4 e variantes reduzem memória, mas exigem teste de qualidade.
- **Observabilidade:** medir tokens/s, TTFT, latência p95/p99, uso de experts, erros, custo por milhão de tokens e saturação de rede.
- **Dados e storage:** RAG, logs, datasets, checkpoints e objetos precisam de storage de alta capacidade e governança.

Camada	Funcao	Ponto de atenção
Compute GPU	Executar experts, atenção e MLPs	VRAM, NVLink, PCIe, FP8/INT8, multi-GPU
Rede	Sincronizar experts e lotes	InfiniBand/Ethernet 100/200/400G, latência, all-to-all
Storage	Model weights, snapshots, datasets, RAG	S3, NFS/Lustre, throughput, versionamento
Serving	Fila, batching, cache, roteamento	Autoscaling, p95/p99, fallback, custo por token
Governanca	Seguranca, auditoria, LGPD, dados	Isolamento por cliente, logs, criptografia, IAM

Implicação para data center
MoE aumenta a importancia de conectividade interna de alta capacidade, energia previsivel, resfriamento adequado, cluster GPU bem desenhado e storage escalavel. Para ofertas de LLMaaS e bare metal GPU, a arquitetura do data center passa a ser parte do produto.

7. Impacto para negocios e produtos de IA

Como transformar MoE em valor comercial

Para o cliente final, a sigla MoE so importa quando melhora resultado, custo, latência ou capacidade. O papel de uma empresa de tecnologia e traduzir a arquitetura para beneficios concretos: automação mais rapida, agentes mais precisos, menos custo de inferência e infraestrutura mais adequada.

Aplicacoes com maior aderência:

- **LLMaaS corporativo:** oferecer modelos via API com controle de custo, isolamento e observabilidade.
- **Agentes de IA:** executar tarefas com ferramentas, memória, RAG, workflows e governança.
- **RAG empresarial:** combinar modelos MoE com bases documentais, contratos, tickets, CRM e dados internos.
- **Vibe coding e automação:** usar modelos fortes em código para acelerar desenvolvimento e integracoes.
- **Atendimento e copilotos:** usar roteamento entre modelos: pequenos para tarefas simples, MoE para tarefas complexas.

Oferta	Mensagem comercial	Entregavel possivel
LLMaaS	IA escalavel com custo por token controlado.	API gerenciada, dashboard, limites, logs, modelos por perfil.
GPU Bare Metal	Infraestrutura dedicada para modelos abertos e MoE.	Clusters H100/H200/MI300, rede, storage, suporte.
RAG Empresarial	Conhecimento interno com segurança e auditoria.	Pipeline de dados, vetorial, S3/NFS, politicas LGPD.
Agentes	Workflows inteligentes com governança.	Agentes por departamento, tool-use, aprovações, relatórios.
Consultoria IA	Escolha técnica entre dense, MoE e modelos fechados.	Assessment, POC, benchmark e plano de produção.

8. Recomendacoes para EnQ Digital

Como usar o tema MoE como autoridade e oferta

A EnQ Digital pode usar o tema “Players LLM x MoE” como conteúdo de autoridade para conectar tres narrativas: tecnologia de IA, infraestrutura de data center e transformação digital. O diferencial esta em explicar a arquitetura de forma executiva e, ao mesmo tempo, mostrar dominio técnico de GPU, rede, storage, cloud privada e LLMaaS.

Plano de conteúdo recomendado:

- **Post 1:** “O que e MoE?” - analogia dos especialistas e router.
- **Post 2:** “Players LLM x MoE” - tabela com confirmados publicamente.
- **Post 3:** “Por que MoE muda o custo da IA?” - parâmetros totais x ativos.
- **Post 4:** “Infra para IA” - GPU, rede, storage, energia e cooling.
- **Post 5:** “Como a EnQ ajuda” - LLMaaS, bare metal, cloud, colocation e conectividade.

Mensagem-chave para vendas:

Proposta de valor
A EnQ Digital conecta IA, cloud, data center e conectividade para empresas que querem transformar modelos de linguagem em produtos reais, com infraestrutura segura, escalavel e observavel.

Checklist para uma POC MoE/LLM:

- Definir caso de uso: RAG, agente, atendimento, código, analise de documentos ou automação.
- Selecionar modelos: fechado, open-weight, dense pequeno, dense grande ou MoE.
- Medir qualidade: acurácia, alucinação, aderência a instrucoes, segurança e benchmark interno.
- Medir operação: latência p95/p99, tokens/s, TTFT, custo por milhão de tokens e uso de GPU.
- Validar governança: LGPD, isolamento, logs, criptografia, IAM e retenção de dados.
- Projetar escala: capacidade de GPU, storage, rede, backup, observabilidade e suporte.

9. Glossario essencial

Termos para usar em reunioes, posts e propostas

Termo	Definicao objetiva
MoE	Mixture of Experts: arquitetura que divide partes do modelo em especialistas e ativa apenas alguns por token.
Router/Gate	Componente que decide quais especialistas processam cada token.
Parametro total	Quantidade total de pesos do modelo, incluindo experts que podem nao ser ativados em cada token.
Parametro ativo	Quantidade de parâmetros usados para processar um token ou entrada em uma passagem.
Top-k	Numero de especialistas escolhidos pelo roteador por token.
Expert parallelism	Distribuicao dos especialistas entre GPUs ou servidores para viabilizar treinamento/inferência.
All-to-all	Padrao de comunicação frequente em MoE quando tokens precisam ser enviados a experts diferentes.
KV cache	Memoria usada para acelerar geração em modelos autoregressivos; cresce com contexto e batch.
Dense model	Modelo em que a maior parte da rede relevante e acionada de forma uniforme a cada token.

10. Fontes e leitura recomendada

Base publica usada para o mapa de players

[1] Shazeer et al. (2017) - Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538.

[2] Fedus, Zoph & Shazeer (2021/2022) - Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961 / JMLR.

[3] Google - Introducing Gemini 1.5, Google's next-generation AI model. Blog oficial do Google.

[4] Mistral AI - Mixtral 8x7B model card e “Mixtral of Experts”. Docs e publicação oficial da Mistral.

[5] DeepSeek - Introducing DeepSeek-V3 e DeepSeek-V3 Technical Report.

[6] Qwen / Alibaba - Qwen3 blog, Qwen3 Technical Report e model cards Qwen3-235B-A22B.

[7] Meta - The Llama 4 herd: natively multimodal AI innovation. Blog oficial Meta AI e model cards Llama 4.

[8] xAI - Open Release of Grok-1 e repositorio xai-org/grok-1.

[9] Databricks - Introducing DBRX: A New State-of-the-Art Open LLM. Blog oficial Databricks.

[10] Snowflake - Snowflake Arctic: enterprise-grade LLM e dense-MoE hybrid architecture.

[11] AI21 - Jamba: Hybrid Transformer-Mamba MoE architecture. AI21 research/blog.

[12] Moonshot AI - Kimi K2 technical report, GitHub/Hugging Face e plataforma Kimi API.

[13] MiniMax - MiniMax-Text-01 / MiniMax-M1 repositorios, model cards e relatórios técnicos.

[14] OpenAI - GPT-4 Technical Report: arquitetura e detalhes de treinamento nao divulgados no relatorio publico.

Nota de verificação

O mapa de players deve ser revisado periodicamente. Fornecedores alteram nomenclatura, disponibilidade, preços, contexto, endpoints e licenças. Para proposta comercial, validar sempre o model card e os termos de uso vigentes do fornecedor.

IA, Cloud e Data Center

Infraestrutura inteligente para empresas que nao podem parar.

Este material foi produzido para apoiar conteúdo educativo, vendas consultivas e conversas técnicas sobre LLMs, MoE e infraestrutura de IA.

EnQ Digital

Cloud | IA | Data Center | Conectividade