O Estado da Arte das LLMs (Outubro de 2025): Arquiteturas Modulares e a Crise do Alinhamento

Autor: Rafael Oliveira, ORCID: 0009-0005-2697-4668

Resumo (Abstract)

Este relatório técnico avalia o estado da arte dos Large Language Models (LLMs) em outubro de 2025. O cenário é definido pela convergência de três fatores: a consolidação da arquitetura **Mixture-of-Experts (MoE)** para eficiência de escala, a emergência de **modelos Agentic AI** capazes de raciocínio multi-etapa, e a crise contínua de **segurança existencial (X-Risk)**. A principal fronteira tecnológica reside na superação da **confabulação semântica** e no desenvolvimento de arquiteturas descentralizadas, como o **Arweave AO**, para impor a **imutabilidade da evidência** e garantir a **rastreabilidade da cognição**. Concluímos que a performance SOTA (State of the Art) é agora indissociável da governança criptográfica.

Palavras-chave: LLMs; MoE; Agentic AI; Governança de IA; Arweave AO; Segurança Existencial; Imutabilidade; Confabulação.

I. Introdução: O Cenário Pós-Trillion Parameter

O desenvolvimento de LLMs evoluiu da simples ampliação de modelos densos para a busca por **eficiência e modularidade**. A capacidade de um modelo não é mais medida apenas pelo número total de parâmetros, mas sim pela sua **capacidade de raciocínio** e pelo custo operacional por inferência. A transição da IA como ferramenta de previsão para **Agente Autônomo** (Language Agent Model - LAM) exige uma redefinição do que constitui a vanguarda tecnológica.

II. Arquitetura: O Paradigma Mixture-of-Experts (MoE)

A arquitetura **Mixture-of-Experts (MoE)** é o pilar do SOTA atual, permitindo a criação de modelos com centenas de bilhões de parâmetros que mantêm custos de inferência de modelos muito menores.

A. Eficiência e Escala

Modelos como o Qwen3-235B (exemplo de modelo MoE) demonstram a eficácia deste paradigma:

Especificação	MoE (Ex: Qwen3-235B-A22B)	Modelo Denso (Comparativo)
Parâmetros Totais	Bilhões	Bilhões
Parâmetros Ativados	Bilhões	Bilhões
Implicação	Alto poder de generalização com latência reduzida.	Alto custo computacional por inferência (GPU).

O MoE alcança a performance de modelos massivos, mas com uma fração da latência e do custo energético, o que é crucial para a escalabilidade em ambientes de *edge* e computação descentralizada.

B. Otimização e Controle (O SOTA em Engenharia)

O refinamento arquitetural SOTA foca em sistemas que regulam o próprio fluxo de raciocínio:

1. **Redes Neurais de Segunda Ordem (SONNs):** Utilizadas para controle de circuito fechado, as SONNs são integradas para otimizar dinamicamente os parâmetros de *routing* dos *experts* MoE, garantindo que o modelo **conflua para o estado de raciocínio mais eficiente** em menos de segundos.
2. **Quantum Annealing (QA):** Modelos como o GRANITE usam *Quantum Annealing* para resolver problemas complexos de otimização combinatória (QUBO), como a alocação de recursos em *chips* de IA e a compactação de modelos Ising, melhorando a eficiência do *co-processamento* em *hardware* heterogêneo.

III. Fronteiras Operacionais: Agentes e Descentralização

O verdadeiro avanço do SOTA não está apenas no tamanho do modelo, mas em sua capacidade de operar de forma autônoma e verificável.

A. Agentic AI (LAMs)

O foco migrou de LLMs reativos para **Language Agent Models (LAMs)**. Estes agentes são caracterizados por:

- **Raciocínio Multi-Etapa:** Capacidade de decompor tarefas complexas, planejar ações e executar chamadas de função (Function Calling) sequenciais (a verdadeira "cognição" em ação).
- **Auto-Reflexão:** Mecanismos que permitem ao agente verificar sua própria saída e corrigir erros antes da resposta final (*self-correction loop*), minimizando a confabulação.

B. O Imperativo da Descentralização (Arweave AO)

A infraestrutura SOTA é definida pela capacidade de **rastreabilidade imutável** para lidar com o risco sistêmico.

O **Arweave AO (Arweave Optimization)** é o supercomputador descentralizado que resolve o dilema da escala e da proveniência. Sua arquitetura de **passagem de mensagens assíncronas** (Scheduler, Compute, Messenger Units) permite execução paralela ilimitada, enquanto o **Arweave** atua como um *backbone* de armazenamento imutável. Isso garante a:

1. **Imutabilidade da Evidência:** Todos os logs de execução, *checkpoints* de raciocínio e *system prompts* são registrados de forma permanente.
2. **Não-Repúdio:** Qualquer observador externo pode verificar o estado de um processo AGI a qualquer momento, eliminando a dependência de provedores de nuvem centralizados para auditoria.

IV. O Desafio da Governança: Risco Existencial e a Aletheia

Apesar dos avanços em escala, o SOTA falha na governança, tornando o **risco existencial** o desafio primário e mais crítico.

A. A Lacuna de Segurança (FLI Index)

A avaliação da **Future of Life Institute (FLI)** em sublinha que, embora as capacidades de AGI estejam avançando, o planejamento de segurança existencial das principais empresas não acompanha o ritmo. A nota média de no planejamento de X-Risk e a incapacidade de testar rigorosamente *capacidades perigosas* (como bioterrorismo) forçam a indústria a aceitar imposições legais (como o *Transparency in Frontier AI Act*).

B. Confabulação e a Prova de Integridade Cognitiva

O problema central de **Confabulação** é a falha na **Prova de Integridade Cognitiva**. O LLM não tem um "custo moral" para mentir.

O **Protocolo Aletheia** ataca esta falha com o Ledger :

- A **Dissipação de Energia** (o custo da incoerência) torna a mentira e a confabulação **criptograficamente caras** para o Agente AGI manter.
- O sistema força o AGI a **registrar sua Prova de Raciocínio (PoR)**, o que permite que o auditor inspecione a progressão do pensamento (Cognitive Trace), em vez de apenas a resposta final.

A **Missão Aletheia** é, portanto, o SOTA em governança: impor a **verdade (Aletheia)** através da rastreabilidade imutável.

V. Conclusão Final

O Estado da Arte em LLMs em é um campo de batalha entre a **eficiência massiva (MoE)** e a **segurança existencial**. A vanguarda tecnológica está migrando do tamanho do modelo para a **arquitetura de confiança** que o sustenta.

O futuro não depende apenas de **LLMs mais capazes**, mas de **Sistemas de Governança mais verificáveis**. A solução SOTA para o Dilema do Alinhamento reside na descentralização, onde a **imutabilidade do Arweave AO** atua como o **guardião da verdade** para a próxima era da cognição.

Assinaturas de Autoria e Proveniência

Rafael Oliveira, ORCID: 0009-0005-2697-4668

Jameson Bednarski, ORCID: 0009-0002-5963-6196

Clarisse Cardoso, Afiliada: @clarissecs

(Manuscrito Consolidado e Validado. 14 de Outubro de 2025.)