DeepSeek V3.2 e NVIDIA DGX Spark: a IA desce à secretária

Publicado em 2025-12-22 11:38:14

BOX DE FACTOS

- **DeepSeek-V3.2-Exp** (29 Set 2025) introduz **DeepSeek Sparse Attention (DSA)**, com ganhos de

directamente no **tool-use** e suporta ferramentas em modos “thinking” e “non-thinking”.˝1˜

- **DeepSeek-V3.2-Speciale** (1 Dez 2025) surge como variante de alto desempenho no Hugging Face.˝2˜
- **NVIDIA DGX Spark** é vendido a **$3,999** no marketplace NVIDIA; usa o **GB10 Grace Blackwell**, 1 PFLOP FP4 e 128GB de memória unificada.˝3˜
- **NVIDIA DGX Station** (GB300 Grace Blackwell Ultra) é anunciada com até **784GB** de memória coerente para treino/inferência de larga escala no desktop.˝4˜

DeepSeek V3.2 e NVIDIA DGX Spark: a IA desce à secretária — e começa a trabalhar

> *Há um momento em que a tecnologia deixa de ser promessa e passa a ser ferramenta: não fala alto, não*

Legenda: a nova geração de "IA local" — do laboratório para o desktop, e do desktop para o negócio.

1) O que mudou no DeepSeek: versões, ambição e pragmatismo

Em Setembro de 2025, o DeepSeek lançou o **V3.2-Exp** como "passo intermédio" rumo à próxima geração. A frase é humilde, mas a engenharia por trás não é: entra em cena o **DeepSeek Sparse Attention (DSA)**, uma forma de tornar o contexto longo menos caro — em memória, em computação e, no fim, em euros na factura.́5˜

Depois, a 1 de Dezembro, surge o **DeepSeek-V3.2** "oficial": o destaque não é apenas performance; é a

inconsistências quando a tarefa exige passos.˝6˜

E aparece ainda a variante **V3.2-Speciale**, sinal claro de maturidade de oferta: quando há "edições" para desempenho, a casa já está a pensar em implementação real, não só em demonstração.˝7˜

2) Porquê o DGX Spark: a máquina que devolve a IA ao controlo local

O **DGX Spark** é a aposta da NVIDIA na "IA local séria": um sistema compacto, com o **GB10 Grace Blackwell**, desenhado para inferência e afinação de modelos grandes com latência mínima e dados sob controlo. No marketplace NVIDIA, o preço de referência surge nos **$3,999**, o que o coloca num patamar curioso: caro como workstation, barato como infra-estrutura de IA.˝8˜

O ponto decisivo não é a vaidade do hardware; é a liberdade operacional: **dados sensíveis não saem**, a latência cai, e o custo deixa de ser um taxímetro de cloud. Para PME, isto é vital: a IA deixa de ser "serviço externo" e passa a ser **capacidade interna**.

A implementação inteligente não começa por "treinar do zero". Começa por **RAG** (recuperação aumentada por geração), regras de segurança, e só depois — quando o retorno o justifica — por afinação.

- **Fase A — RAG**: indexar documentos (PDF, DOCX, e-mails, catálogos, tabelas) e ligar o modelo ao repositório; perguntas passam a ser respondidas com base em fontes internas e citações.
- **Fase B — Agentes**: aproveitar o "tool-use" do V3.2 para chamar funções do sistema: procurar preços, gerar orçamentos, validar stock, escrever propostas, abrir tickets.
- **Fase C — Afinação**: apenas depois de medir ganhos reais (tempo poupado, erros reduzidos, conversões aumentadas). Aqui o DGX Spark entra como acelerador do ciclo: testar, ajustar, repetir — localmente.

O DeepSeek está a trabalhar explicitamente a eficiência e o contexto longo (DSA), o que casa com RAG e documentos extensos. E, do lado do Spark, a promessa é precisamente essa: correr, afinar e servir modelos no local, com stack NVIDIA pré-alinhada.˝9˜

Quando o projecto deixa de ser "um sistema" e passa a ser "uma fábrica" (vários modelos, vários serviços, múltiplos utilizadores, maior carga), entra a lógica da **DGX Station**. A NVIDIA descreve-a como a primeira estação com o **GB300 Grace Blackwell Ultra** e até **784GB** de memória coerente, orientada a treino e inferência de grande escala no desktop.˝10˘

Em linguagem de negócio: o Spark é a lança; a Station é o aríete. Um abre caminho. O outro sustenta muralhas.

5) Conclusão: a IA deixa de ser "nuvem distante" e passa a ser "motor interno"

O DeepSeek V3.2 está a empurrar a IA para um território mais operativo — agentes, ferramentas, contexto longo, eficiência.˝11˘ E a NVIDIA, com o DGX Spark e a Station, está a empurrar o hardware na direcção oposta à dependência: trazer a capacidade para perto dos dados, perto do programador, perto do negócio.˝12˘

O que nasce aqui é simples e poderoso: uma IA que não vive de promessas, vive de rotina — responde, calcula, propõe, valida, aprende com o ciclo da empresa. E isso, num país que se distrai com ruído, é uma revolução silenciosa.

armazém, ao balcão, ao servidor da PME. E quando a IA chega a esse lugar — deixa de ser moda. Passa a ser infra-estrutura.

Artigo de : Francisco Gonçalves

Softelabs • Tecnologias IT e AI

Nota de co-autoria: texto desenvolvido em colaboração editorial com Augustus.

[leia]