O desafio de implementar um modelo LLM para a Força Aérea Brasileira

1. Odair Oliveira de Sá, Luís Alberto Vieira Dias
Instituto Tecnológico de Aeronáutica (ITA) - São José dos Campos/SP - Brasil

Insti

RESUMO - Este artigo analisa os desafios multidimensionais para a implementação de modelos de linguagem de grande escala (LLM) na Força Aérea Brasileira (FAB). A pesquisa aborda a complexidade da infraestrutura computacional necessária, com foco em processamento paralelo e GPUs, e a imperativa soberania tecnológica para mitigar dependências externas.

Explora-se a integração de fontes doutrinárias (DCA 1-1) e dados de IVR para fomentar a superioridade de informação, mantendo rigorosos controles de segurança cibernética e mitigação de alucinações que possam comprometer a cadeia decisória.

Conclui-se que a IA deve atuar como um multiplicador de força no apoio à decisão, mantendo o ser humano como agente central do Comando e Controle (C2).

I. INTRODUÇÃO

O desafio de criar um modelo de *Large Language Model* (LLM) aplicável à Força Aérea Brasileira (FAB) envolve uma complexidade que perpassa desde a infraestrutura tecnológica necessária até as exigências doutrinárias e de segurança da informação.

O desafio supracitado envolve questões relacionadas com infraestrutura e eficiência computacional, custos e dependência tecnológica, segurança e domínio cibernético, recursos humanos e capacitação, integração doutrinária e interoperabilidade

Em relação à infraestrutura e eficiência computacional, há que se considerar o hardware necessário. Com as arquiteturas de modelos modernos, como o *Transformer,* exige uma otimização matemática direta [5] com as capacidades físicas de GPUs modernas. A eliminação de loops recorrentes para permitir o processamento em paralelo é fundamental para escalar o desempenho, o que demanda um parque tecnológico robusto e atualizado.

Já com relação ao tratamento de dados massivos, a FAB já reconhece a necessidade de integrar dados e análises em tempo real, especialmente com o uso de inteligência artificial para análise de dados massivos e construção de bancos de dados para processos de *targeting* e consciência situacional.

Os custos, por sua vez, estão relacionados aos altos investimentos e dependência tecnológica para serem superados. Altos investimentos porque requer o desenvolvimento e a manutenção de sistemas de ponta, como a IA, envolvem altos custos não apenas na aquisição de equipamentos e sistemas, mas também no tempo e esforço para formar recursos humanos especializados.

No quesito soberania tecnológica, a FAB reconhece uma dependência de tecnologia, uma vez que o país nem sempre tem a capacidade de gerar ou produzir internamente tudo o que necessita para atender às suas demandas de meios de última geração.

Na questão da segurança e domínio cibernético, sabe-se que LLMs operam no domínio informacional e cibernético. O avanço tecnológico traz o paradoxo de gerar maior dependência e, consequentemente, maiores vulnerabilidades às ações cibernéticas. É consenso absoluto que é imperativo proteger os ativos informacionais contra ataques, explorações e vazamentos, e por isso não pode ser um modelo aberto para uso fora do domínio tecnológico.

Na atividade de inteligência militar, por exemplo, a disseminação de um conhecimento falseado ou impreciso (que pode ocorrer em alucinações de modelos de IA) pode comprometer a cadeia decisória do Comando. O controle da informação significa coletar, processar e proteger dados, negando ao adversário a mesma possibilidade.

Nos aspectos de recursos humanos e capacitação é preciso formar uma massa crítica, pois um dos maiores desafios é a manutenção de pessoal qualificado. Movimentações de efetivo, cursos de carreira e diversos outros fatores que incluem a perda de capital intelectual para corporações privadas devido a maiores atrativos financeiros ou contratuais, pode gerar obsolescência no desenvolvimento de sistemas nacionais.

Por outro lado, a introdução de novas capacidades tecnológicas determina a necessidade de requalificação de efetivo especializado da FAB e novos métodos de treinamento, o que pode limitar o ritmo de mudança que a Força consegue aceitar.

Por fim, na questão da integração doutrinária e interoperabilidade, a Doutrina Básica da FAB [1] preconiza que a própria doutrina deve ser dinâmica para acompanhar o desenvolvimento tecnológico, que hoje inclui o emprego de inteligência artificial e guerra informacional.

Por tudo que foi apresentado até o momento, qualquer modelo deve ser capaz de operar de forma integrada (interoperabilidade) entre as diversas Forças (Marinha, Exército e Aeronáutica) e outros órgãos governamentais, garantindo que a informação flua com rapidez e precisão na cadeia de comando.

II. METODOLOGIA

A pesquisa seguiu uma abordagem qualitativa e exploratória, fundamentada em uma revisão bibliográfica e documental da Doutrina Básica da FAB (DCA 1-1) [1] e de documentos estratégicos como o Força Aérea 100 [2].

O desenvolvimento metodológico dividiu-se em três etapas abaixo discriminadas:

Levantamento de requisitos técnicos com análise da arquitetura Transformer e da necessidade de otimização de hardware soberano (HPC) para processamento em larga escala.

Integração de base de conhecimento com experimento usando COLAB e a proposição de um método de seleção e curadoria de dados operacionais e doutrinários, sob supervisão da atividade de inteligência, para garantir a integridade e autenticidade da informação.

Apoio de uma proposta de um *framework* de Governança [3] com sugestões alinhadas aos padrões de segurança operacional (*safety/security*) do DECEA, para gerenciar o ciclo de vida dos modelos, desde o treinamento em ambiente controlado (*sandbox*) até a operação com parâmetros congelados.

III. COMO IMPLEMENTAR

Para criar uma base de conhecimento para um modelo de LLM que seja segura e adequada às necessidades de uma Força Aérea moderna, é necessário integrar os princípios de superioridade de informação, gestão do conhecimento e segurança da informação e comunicações (SIC) descritos na doutrina militar, além das normas e procedimentos nos níveis tático, operacional e estratégico.

Para treinamento de um modelo LLM, em primeiro lugar é necessário a seleção e integração de fontes doutrinárias e operacionais, de preferência com a supervisão de pessoal de inteligência para garantir que as fontes de dados não estejam comprometidas de alguma forma. A base de conhecimento deve ser alimentada por um conjunto de documentos e dados que reflitam a realidade e a evolução da Força incluindo os documentos de alto nível como a Doutrina Básica [1], a Concepção Estratégica [2] e as diretrizes de política e estratégia nacional de Defesa, assim como também a Política e Estratégia nacional de Defesa [3] e das concepções de emprego das demais forças (Marinha e Exércitos Brasileiro), visto que se deve pensar em conjunto desde a concepção.

Dados de IVR (Inteligência, Vigilância e Reconhecimento) devem integrar dados massivos provenientes de sensores em plataformas aéreas, espaciais e remotamente pilotadas para prover consciência situacional em tempo real também devem constar na base de dados.

Ideal é incorporar relatórios de análises pós-ação (*debriefing*) de exercícios e operações, que são fundamentais para a evolução dinâmica da doutrina.

A base deve ser capaz de fundir dados próprios com bases de dados das demais forças armadas e órgãos governamentais para permitir operações multidomínio (MDO) e uma consciência situacional conjunta visando garantir o máximo de capacidade de interoperabilidade.

O modelo deve atuar na interseção de três domínios para gerar superioridade de decisão: Domínio Físico (plataformas e infraestruturas); Domínio da Informação onde a informação é criada, manipulada e protegida, permitindo o fluxo das intenções do comandante; e o domínio cognitivo que é relacionado à percepção, crenças e tomada de decisão. O LLM deve ajudar o comandante a entender e processar a realidade de forma mais rápida e precisa que o adversário..

IV. EXPERIMENTO COM MODELO DE IA

A quantização 4-bit - via QLoRA (*Quantized Low-Rank Adaptation*) - técnica de *fine-tuning* para LLMs que reduz drasticamente o consumo de memória de GPU [13] - não é apenas uma "técnica de economia de memória". Ela é o divisor de águas que torna a IA um recurso acessível em hardware comum.

A quantização de 4 bits, mencionado acima, refere-se ao processo de reduzir a precisão dos pesos e ativações do modelo para apenas 4 bits. o que significa uma queda drástica em relação aos formatos típicos de ponto flutuante de 16 ou 32 bits. O pré-treinamento com 4 bits exige atenção porque os gradientes e as atualizações devem ser tratados com muito cuidado para preservar a precisão e, ao mesmo tempo, melhorar a velocidade geral do treinamento. Técnicas especializadas são necessárias para manter a eficácia enquanto se mapeia tensores de alta precisão para um conjunto muito menor de valores quantizados [10].

A NVIDIA sugere em seu site [10] que sejam adotadas as técnicas de pré-treinamento indicadas para manter a acurácia do modelo e a convergência estável.

Foram realizados alguns testes usando COLAB, e foi possível aumentar a eficiência de memória (VRAM) com redução de ~70% a 80% do consumo. Um modelo de 7B que ocupava ~14GB em FP16, passa a ocupar de ~3.5GB a 4GB. O que permite rodar o Qwen2.5-7B perfeitamente na GPU T4 do Colab (15 GB), sobrando espaço para o contexto (sequências longas de doutrina) e gradientes de treino.

V. CONCLUSÕES

A implementação de modelos LLM na FAB não representa apenas uma evolução tecnológica, mas uma transformação operacional estratégica necessária para lidar com o ambiente de guerra informacional moderno, tendo em vista que as operações militares estão evoluindo rapidamente das operações conjuntas entre forças para operações multi-domínio, como tem sido visto na Ucrânia.

Conclui-se que, para o êxito dessa implementação, é vital para superar a dependência de tecnologias estrangeiras sensíveis, o investimento em soberania tecnológica e a busca de formas de retenção de capital intelectual qualificado.

Embora os modelos generativos ofereçam ganhos substanciais na análise de dados massivos e na redução do ciclo OODA, a doutrina reforça que a IA deve atuar estritamente como uma ferramenta de apoio no domínio cognitivo, preservando a responsabilidade moral e o julgamento ético do comandante.

Portanto, a adoção de LLMs deve ser pautada por uma governança rigorosa, garantindo que as inovações tecnológicas incrementem a operacionalidade sem comprometer a segurança operacional e a defesa do País.

A colaboração Homem-Máquina (*Human-Machine Teaming*) com o uso de ferramentas de suporte à decisão aprimoradas por IA, com garantia que os seres humanos permaneçam no circuito (*in-the-loop*) para garantir decisões e resultados responsáveis é um dos requisitos para o sucesso das operações multidomínio depende diretamente de obter e manter a vantagem de decisão (*decision advantage*) sobre o adversário

Por fim, há que se ter o "Homem como agente central" e a IA deve ser vista estritamente como um suporte à decisão, preservando o julgamento ético e a responsabilidade moral do comandante, conforme preconiza a doutrina de Comando e Controle.

REFERÊNCIAS

[1] BRASIL. Comando da Aeronáutica. Estado-Maior da Aeronáutica. DCA 1-1: Doutrina Básica da Força Aérea Brasileira. Brasília, DF: EMAER, 2024.
[2] BRASIL. Comando da Aeronáutica. Estado-Maior da Aeronáutica. DCA 11-45: Concepção Estratégica Força Aérea 100. Brasília, DF: EMAER, 2018.
[3] BRASIL. Política Nacional de Defesa, a Estratégia Nacional de Defesa e o Livro Branco de Defesa Nacional. Decreto Nº 12.725, DE 18 DE NOVEMBRO DE 2025
[4] BRASIL. Ministério da Defesa. MD31-M-03: Doutrina para o Sistema Militar de Comando e Controle. 3. ed. Brasília, DF: Ministério da Defesa, 2015.
[5] VASWANI, Ashish et al. *Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[6] DOUHET, Giulio. O Domínio do Ar. Tradução da Escola de Aperfeiçoamento de Oficiais da Aeronáutica. Rio de Janeiro: Instituto Histórico-Cultural da Aeronáutica, 1988.
[7] BRASIL. Comando da Aeronáutica. Estado-Maior da Aeronáutica. Seminário de Governança de IA. Apresentação Maj Eng ISRAEL. Subdepartamento Técnico DECEA. Slides 9-20, 26-36.2026
[8] Disponível em: <https://www.act.nato.int/activities/multi-domain-operations/ >, obtido em Junho de 2026
[9] OTAN. NATO Multi-domain Operations Conference. Ankara, Turkiye. May 2025.
[10] Quantização 4-bits. Disponível em < https://developer-nvidia-com.translate.goog/blog/nvfp4-trains-with-precision-of-16-bit-and-speed-and-efficiency-of-4-bit/?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=tc > obtido em junho de 2026
[11] Tecnologia de defesa disruptiva emergente em operações de vários domínios. Disponível em < https://www.magaero.com/emerging-disruptive-defense-technology-in-multi-domain-operations/ > obtida em set 2026
[12] Monteiro, P.S (Coord) (2025). As áreas funcionais do Exército 2025: desafios do novo ambiente operacional - Impacto da Inteligência Artificial. IUM Atualidade, 52. Instituto Universitário Militar. Pag 140.
[13] Fine_tunning llama2 em hardware de consumo: Um tutorial QloRA completo. Disponível em < https://www.datacamp.com/tutorial/fine-tuning-llama-2?utm_aid=203195303648&utm_loc=9100207-&utm_mtd=-c&utm_kw=&gad_source=1&gad_campaignid=23950607752 > Obtido em SET 2026