****

**UNIVERSIDADE DE BRASÍLIA**

**FACULDADE DE EDUCAÇÃO**

**PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO**

**MESTRADO PROFISSIONAL – POLÍTICAS PÚBLICAS E GESTÃO DA**

**EDUCAÇÃO**

HÉLIO CRAVEIRO PESSOA JÚNIOR

**CROSSDEBATE: UM ORQUESTRADOR DO AJUSTE FINO MULTIAGENTE A PARTIR DE COMPENSAÇÕES ENTRE CARGA COGNITIVA E COMPUTACIONAL**

**Brasília, DF**

**2025**

HÉLIO CRAVEIRO PESSOA JÚNIOR

**CROSSDEBATE: UM ORQUESTRADOR DO AJUSTE FINO MULTIAGENTE A PARTIR DE COMPENSAÇÕES ENTRE CARGA COGNITIVA E COMPUTACIONAL**

Dissertação de Mestrado apresentada ao Programa

de Pós-Graduação em Educação – modalidade

profissional da Universidade de Brasília - UnB

como requisito parcial para obtenção do grau de

Mestre em Educação na linha de pesquisa de

Políticas Públicas e Gestão da Educação.

**Orientador:** Professor Doutor Remi Castioni

Brasília, DF

2025

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Ficha catalográfica elaborada automaticamente,  com os dados fornecidos pelo(a) autor(a) |  |
|  | p475c | Pessoa Júnior, Hélio Craveiro  CROSSDEBATE: UM ORQUESTRADOR DO AJUSTE FINO MULTIAGENTE A PARTIR DE COMPENSAÇÕES ENTRE CARGA COGNITIVA E COMPUTACIONAL / Hélio Craveiro Pessoa Júnior; orientador Remi Castioni. Brasília, 2025.  204 p.  Dissertação (Mestrado Profissional em Educação) Universidade de Brasília, 2025.  1. Fluxos de Trabalho de Análise de Dados. 2. LLMs agenticamente relacionados. 3. Computação Neurofisiológica. 4. Eficiência Computacional. 5. Ajuste Fino de Sistema Generativo/Deterministico. I. Castioni, Remi , orient. II. Título. |  |

HÉLIO CRAVEIRO PESSOA JÚNIOR

**CROSSDEBATE: UM ORQUESTRADOR DO AJUSTE FINO MULTIAGENTE A PARTIR DE COMPENSAÇÕES ENTRE CARGA COGNITIVA E COMPUTACIONAL**

Dissertação de Mestrado apresentada ao Programa

de Pós-Graduação em Educação – modalidade

profissional da Universidade de Brasília - UnB

como requisito parcial para obtenção do grau de

Mestre em Educação na linha de pesquisa de

Políticas Públicas e Gestão da Educação.

**Orientador:** Professor Doutor Remi Castioni

Defendida e aprovada em: 12 de junho de 2025.

COMISSÃO EXAMINADORA:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Remi Castioni - Universidade de Brasília (presidente)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Bernardo Kipnis - Universidade de Brasília (avaliador interno)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Pedro Carvalho Brom - Instituto Federal de Brasília - IFB (avaliador externo)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dra. Girlene Ribeiro de Jesus - Universidade de Brasília (suplente)

Brasília, DF

2025

**RESUMO**

Este estudo investiga a interdependência dinâmica entre a Carga Cognitiva (CL) do operador humano e a Carga Computacional (CompL) em sistemas de inteligência artificial local, num contexto de democratização tecnológica. O foco está no ajuste fino via QLoRA e na orquestração de fluxos multiagente (sequenciais e paralelos), empregando modelos LLMs GGUF (Q4 vs. Q8) em hardware acessível. Foi implementada uma plataforma React/FastAPI, com um enxame de 60 modelos, fundamentada na Teoria da Carga Cognitiva (Sweller, 2011) e nos princípios ergonômicos da ISO 9241-210:2019. A CL foi avaliada por meio de instrumentos psicométricos (NASA-TLX Adaptado, SEQ) combinados com monitoramento neurofisiológico contínuo e sincronizado (EEG, rastreamento ocular/pupilometria, PPG e SpO2). Já a CompL foi medida por métricas de desempenho de hardware (RAM, VRAM, uso de GPU, consumo energético e temperatura). O experimento manipulou sistematicamente a quantização GGUF, a complexidade dos fluxos e o nível de expertise dos operadores. Os resultados confirmaram que a quantização Q4 reduziu CompL, mas aumentou significativamente a CL devido à degradação da qualidade e comportamentos imprevisíveis (H1). A complexidade crescente dos fluxos amplificou ambas as cargas exponencialmente (H2). A relação entre CL e CompL mostrou-se não linear (formato U/J), revelando um “ponto ideal” em Q8 que minimiza a CL (H3). A expertise do operador atua como moderador, mitigando o impacto da CL em especialistas (H4). Esses achados reforçam a necessidade urgente de plataformas de AI locais co-adaptativas, capazes de equilibrar dinamicamente CL e CompL para promover interações ergonômicas sustentáveis. A plataforma e os resultados estão disponíveis em <http://crossdebate.com>.

**Palavras-chave:** Fluxos de Trabalho de Análise de Dados; LLMs agenticamente relacionados; Computação Neurofisiológica; Eficiência Computacional; Ajuste Fino de Sistema Generativo/Deterministico.

**ABSTRACT**

This study investigates the critical and dynamic interdependence between human operator Cognitive Load (CL) and system Computational Load (CompL), a central challenge in the democratized AI era, focusing on local fine-tuning (QLoRA) and orchestration of multiagent workflows (sequential, parallel) with GGUF LLMs (Q4 vs. Q8) on consumer hardware. By designing a platform (React/FastAPI) with a swarm of 60 models, the research employed a transdisciplinary mixed-methods methodology grounded in Cognitive Load Theory (Sweller, 2011) and ergonomic principles from ISO 9241-210:2019. CL was evaluated through a robust crossmodal approach, combining psychometric instruments (Adapted NASA-TLX, SEQ) with continuous neurophysiological monitoring synchronized via Lab Streaming Layer (LSL), including eye tracking/pupillometry (POG), electroencephalography (EEG), photoplethysmography (PPG) for heart rate variability, and peripheral oxygen saturation (SpO2). Simultaneously, CompL was quantified by hardware metrics (RAM, VRAM, GPU usage, energy, thermal temperature). The experimental design systematically manipulated GGUF quantization, multiagent flow complexity, and operator expertise. Results empirically confirmed that aggressive quantization (Q4) reduced CompL but substantially increased CL through quality degradation and behavioral unpredictability (H1). Increasing flow complexity elevated both loads exponentially (H2). Crucially, the CL-CompL relationship revealed itself as non-linear (U/J-shaped), indicating a "sweet spot" of moderate CompL/quality (Q8) that minimizes CL (H3). Expertise acted as a significant moderator, dampening CL increases for specialists (H4). The findings underscore the urgent need for co-adaptive local AI platforms that actively manage the CL-CompL balance in real-time, targeting the optimal zone of sustainable ergonomic interaction. The platform and results are available at <http://crossdebate.com>.

**Keywords:** Data Analysis Workflows; Agentic LLM Workflows; Neurophysiological Computing; Computational Efficiency; Generative/Deterministic System Fine-Tuning.

**LISTA DE ABREVIATURAS E SIGLAS**

A2A – Agent-to-Agent (Comunicação entre Agentes)

AI – Artificial Intelligence (Inteligência Artificial)

AOI – Area of Interest (Área de Interesse)

API – Application Programming Interface (Interface de Programação de Aplicações)

AutoPDL – Automatic Prompt Definition Language

CL – Carga Cognitiva

CLT – Cognitive Load Theory (Teoria da Carga Cognitiva)

CompL – Carga Computacional

CPU – Central Processing Unit (Unidade Central de Processamento)

ECL – Extraneous Cognitive Load (Carga Cognitiva Extrínseca)

EEG – Eletroencefalografia

EMM – Estimated Marginal Means (Médias Marginais Estimadas)

FDR – False Discovery Rate (Taxa de Falsas Descobertas)

FFT – Full Fine-Tuning (Ajuste Fino Completo)

FMOps – Foundation Model Operations

FWER – Family-Wise Error Rate (Taxa de Erro Familiar)

GAMM – Generalized Additive Mixed Models (Modelos Aditivos Generalizados Mistos)

GCL – Germane Cognitive Load (Carga Cognitiva Germânica)

GGUF – GPT-Generated Unified Format

GLMM – Generalized Linear Mixed Models (Modelos Lineares Generalizados Mistos)

GPT – Generative Pre-trained Transformer

GPU – Graphics Processing Unit (Unidade de Processamento Gráfico)

HCI – Human-Computer Interaction (Interação Humano-Computador)

HF – High Frequency (Alta Frequência - em HRV)

HITL – Human-in-the-Loop (Humano no Ciclo)

HoT – Hypergraph of Thought

HRV – Heart Rate Variability (Variabilidade da Frequência Cardíaca)

HSD – Honestly Significant Difference (Diferença Honestamente Significativa - de Tukey)

IA – Inteligência Artificial

IBI – Intervalos Inter-Batimentos

ICA – Análise de Componentes Independentes

ICL – Intrinsic Cognitive Load (Carga Cognitiva Intrínseca)

ICC – Intraclass Correlation Coefficient (Coeficiente de Correlação Intraclasse)

IDE – Integrated Development Environment (Ambiente de Desenvolvimento Integrado)

IIQ – Intervalo Interquartil

ISO – International Organization for Standardization

KDE – Kernel Density Estimation (Estimativa de Densidade Kernel)

LF – Low Frequency (Baixa Frequência - em HRV)

LLM – Large Language Model (Grande Modelo de Linguagem)

LLaMA – Large Language Model Meta AI

LoRA – Low-Rank Adaptation

LRT – Likelihood Ratio Test (Teste de Razão de Verossimilhança)

LSL – LabStreamingLayer

MARL – Multi-Agent Reinforcement Learning

MCAR – Missing Completely At Random

MCP – Model Context Protocol

MICE – Multiple Imputation by Chained Equations

ML – Maximum Likelihood (Máxima Verossimilhança)

NASA-TLX – NASA Task Load Index

NL Web – Natural Language Web

NVMe – Non-Volatile Memory Express

OOM – Out Of Memory (Falta de Memória)

OR – Odds Ratio (Razão de Chances)

PDL – Prompt Definition Language

PEFT – Parameter-Efficient Fine-Tuning

POG – Point of Gaze (Ponto de Olhar - Rastreamento Ocular)

PPG – Fotopletismografia

PSD – Densidade Espectral de Potência

PTQ – Post-Training Quantization (Quantização Pós-Treinamento)

Q4 – Quantização de 4 bits

Q8 – Quantização de 8 bits

QLoRA – Quantized Low-Rank Adaptation

Q-Q – Quantil-Quantil

RAG – Retrieval-Augmented Generation

RAM – Random Access Memory (Memória de Acesso Aleatório)

REML – Restricted Maximum Likelihood (Máxima Verossimilhança Restrita)

RMSSD – Root Mean Square of Successive Differences (em HRV)

RMSEA – Root Mean Square Error of Approximation

SDNN – Standard Deviation of NN intervals (em HRV)

SEQ – Single Ease Question

SFT – Supervised Fine-Tuning (Ajuste Fino Supervisionado)

SO – Sistema Operacional

SpO2 – Saturação Periférica de Oxigênio

SSD – Solid State Drive (Unidade de Estado Sólido)

STV – Single Transferable Vote (Voto Único Transferível)

SVM – Support Vector Machines (Máquinas de Vetores de Suporte)

TPS – Tokens por Segundo

UI – User Interface (Interface do Usuário)

UX – User Experience (Experiência do Usuário)

VAR – Vetores Autorregressivos

VD – Variável Dependente

VI – Variável Independente

VRAM – Video Random Access Memory (Memória de Acesso Aleatório de Vídeo)

**LISTA DE SÍMBOLOS**

α (alfa) – Nível alfa (probabilidade de erro Tipo I); hiperparâmetro em LoRA

β (beta) – Probabilidade de erro Tipo II (em análise de poder, 1-β é o poder); coeficiente de regressão

∆W – Mudança nos pesos (em LoRA)

η²p – Eta-quadrado parcial (tamanho do efeito em ANOVA/GLMM)

κ (kappa) – Kappa de Cohen (concordância interavaliadores)

µV²/Hz – Microvolts quadrados por Hertz (unidade de potência espectral EEG)

ρ (rho) – Coeficiente de correlação de Spearman; correlação entre medidas repetidas

Σ (sigma maiúsculo) – Somatório

σ (sigma minúsculo) – Desvio padrão; erro

τ (tau) – Coeficiente de correlação tau de Kendall; constante de tempo em curvas de aprendizado

CI – Confidence Interval (Intervalo de Confiança)

d – d de Cohen (tamanho do efeito para diferença entre médias)

df – Graus de liberdade (degrees of freedom)

F – Estatística F (em ANOVA/GLMM)

GB – Gigabytes

Hz – Hertz

kWh – Quilowatt-hora

ms – Milissegundos

N – Número total de participantes na amostra

n – Número de participantes em um subgrupo ou número de observações

p – Valor-p (nível de significância estatística)

r – Coeficiente de correlação de Pearson

W – Watts

± – Mais ou menos (indicando variabilidade ou margem de erro)

√ – Raiz quadrada

≤ – Menor ou igual a

≥ – Maior ou igual a

**SUMÁRIO**

|  |  |
| --- | --- |
| **INTRODUÇÃO** | 13 |
| **REVISÃO DA LITERATURA** | 31 |
| 2.1 Grandes Modelos de Linguagem e a Mudança para a Execução Local | 33 |
| 2.2 Ajuste Fino de Modelos Quantizados (PEFT/QLoRA) | 38 |
| 2.3 Teoria da Carga Cognitiva (CLT) na Interação Humano-Computador | 43 |
| 2.4 Medindo a Carga Cognitiva | 45 |
| 2.5 Métricas e Medição da Carga Computacional (CompL) | 50 |
| 2.6 Desafios da Orquestração de Enxames de Agentes LLM em Ambientes de Recursos Limitados: Perspectivas da Carga Cognitiva e Computacional | 52 |
| 2.7 A Interdependência da Carga Cognitiva e Computacional: Lacunas na Literatura | 56 |
| 2.8 Conclusão da Revisão da Literatura | 58 |
| **METODOLOGIA** | 59 |
| 3.1 Abordagem e Desenho da Pesquisa | 60 |
| 3.2 Quadro Conceitual | 62 |
| 3.3 Variáveis | 64 |
| 3.4 Participantes | 70 |
| 3.5 Materiais e Aparatos | 74 |
| 3.6 Procedimento Experimental | 77 |
| 3.7 Medição de Dados | 80 |
| 3.8 Rigor e Validade | 81 |
| **ANÁLISE DE DADOS** | 83 |
| 4.1 Visão Geral da Estratégia Analítica | 84 |
| 4.2 Preparação e Limpeza de Dados | 87 |
| 4.3 Estatísticas Descritivas | 90 |
| 4.4 Estatísticas Inferenciais (Teste de Hipóteses via GLMMs) | 91 |
| 4.5 Análise de Correlação e Regressão (Interdependência - RQ3 e H3) | 95 |
| 4.6 Análise de Dados Qualitativos (Análise Temática) | 96 |
| 4.7 Integração de Métodos Mistos | 97 |
| **RESULTADOS** | 98 |
| 5.1 H1: Efeitos do Nível de Quantização GGUF na CL e CompL | 100 |
| 5.1.1 Impacto na Carga Computacional (CompL) | 100 |
| 5.1.2 Impacto na Carga Cognitiva (CL) | 105 |
| 5.1.3 Impacto no Resultado da Tarefa | 114 |
| 5.1.4 Conclusão Parcial para H1 | 117 |
| 5.2 H2: Efeitos da Complexidade do Fluxo de Trabalho Multiagente na CL e CompL | 117 |
| 5.2.1 Impacto na Carga Computacional (CompL) | 117 |
| 5.2.2 Impacto na Carga Cognitiva (CL) | 119 |
| 5.2.3 Impacto no Resultado da Tarefa | 121 |
| 5.2.4 Conclusão Parcial para H2 | 121 |
| 5.3 H4: Papel Moderador da Expertise na Relação CL-CompL | 122 |
| 5.3.1 Conclusão Parcial para H4 | 124 |
| 5.4 RQ3/H3: Natureza da Relação CL-CompL (Não Linearidade) | 124 |
| 5.4.1 Conclusão Parcial para RQ3/H3 | 126 |
| 5.5 Síntese dos Achados Gerais da Seção de Resultados | 127 |
| **DISCUSSÃO** | 128 |
| 6.1 Visão Geral: Navegando na Interdependência Cognitivo-Computacional em Sistemas de AI Locais | 129 |
| 6.2 O Custo Cognitivo da Economia Computacional: Interpretando o Balanço da Quantização GGUF (H1) | 133 |
| 6.3 O Fardo da Orquestração: Complexidade do Fluxo de Trabalho Multiagente, CL e CompL (H2) | 139 |
| 6.4 O Ponto Ideal: Aprofundando a Não Linearidade na Relação CL-CompL (H3) | 149 |
| 6.5 O Papel de Amortecedor da Expertise (H4) | 153 |
| 6.6 Sintetizando a Interação: Rumo a Sistemas Co-Adaptativos Humano-IA | 157 |
| **CONCLUSÃO** | 161 |
| **REFERÊNCIAS** | 165 |
| **APÊNDICE I** - NASA-TLX Adaptado | 198 |
| **APÊNDICE II** - Glossário de Termos Técnicos | 202 |

**1. INTRODUÇÃO**

A inteligência artificial (IA) contemporânea encontra-se em um ponto de inflexão, impulsionada por avanços exponenciais na capacidade e acessibilidade dos Grandes Modelos de Linguagem (LLMs). Arquiteturas seminais, como as famílias GPT (OpenAI, 2025; Kumar et al., 2025), LLaMA (Touvron et al., 2023; Kim et al., 2025; Wang et al., 2025a), Mistral (Jiang et al., 2023) e Claude (Anthropic, 2025a, 2025b), demonstraram uma proficiência sem precedentes em tarefas que vão desde a geração de linguagem natural e tradução automática até formas complexas de raciocínio, planejamento e resolução de problemas em múltiplos domínios (Zhao e Yaun, 2025; Chang et al., 2023; Liu et al., 2025). Recentemente, Khramov et al. (2025) destacam que a integração de IA generativa com sistemas de prova interativa, como o Rocq, amplia ainda mais esse espectro, permitindo automação em tarefas tradicionalmente consideradas altamente especializadas, como verificação formal de software. Os autores mostram que abordagens baseadas em recuperação de similaridade e sistemas multiagente — especialmente com debates entre agentes para planejamento de provas — elevam substancialmente a eficácia e a robustez dessas soluções. Isso evidencia que o avanço dos LLMs não apenas expande o alcance da IA em tarefas linguísticas, mas também transforma profundamente áreas técnicas, promovendo automação confiável e colaborativa em domínios críticos como a engenharia de software formal.

Este progresso tem sido, em parte, atribuído a leis de escala empíricas que correlacionam o desempenho aprimorado com o aumento exponencial no número de parâmetros do modelo e no volume de dados de pré-treinamento (Kaplan et al., 2020; Hoffmann et al., 2022). Capacidades não explicitamente presentes em modelos menores, mas que surgem quando os modelos ultrapassam certos limiares de escala, as "habilidades emergentes" resultantes (Wei et al., 2022b; Ouyang et al., 2022) sugeriram um caminho, embora computacionalmente oneroso, em direção a sistemas de AI com capacidades cognitivas cada vez mais generalizadas, aproximando as máquinas da cognição de nível humano em domínios restritos.

Contudo, este promissor avanço em direção à cognição artificial generalizada revela um paradoxo fundamental: enquanto as capacidades emergentes expandem exponencialmente o potencial transformador dos LLMs, elas simultaneamente intensificam de forma dramática as demandas computacionais, criando uma tensão crescente entre inovação tecnológica e viabilidade prática que define os contornos atuais do desenvolvimento em AI.

Modelos mais recentes, como o DeepSeek-R1-Distill-Qwen-7B, exemplificam essa tendência, incorporando etapas explícitas de raciocínio para melhorar a qualidade das respostas em tarefas complexas (Marjanovic et al., 2025), um processo que, embora vigoroso, adiciona novas camadas de complexidade computacional e interpretativa. Esta abordagem de raciocínio expandido durante o teste conecta-se com desenvolvimentos em compute scaling durante inferência, como demonstrado por Lin et al. (2025) em seu trabalho sobre Sleep-time Compute, que explora como recursos computacionais adicionais em tempo de teste podem ser aproveitados para melhorar o desempenho de modelos sem retreinamento, oferecendo uma perspectiva complementar sobre como balancear qualidade de resposta e demanda computacional em sistemas de AI local com recursos limitados. Em linha com essa discussão, Cremer (2025) destaca como LLMs podem ser integrados a fluxos de trabalho complexos, automatizando tarefas tradicionalmente realizadas por especialistas humanos, como a personalização de contratos de eletricidade em larga escala. O autor mostra que, ao permitir interações naturais entre usuários e sistemas técnicos via linguagem natural, LLMs não apenas aumentam a eficiência e a personalização, mas também reduzem drasticamente os custos de transação e o tempo de resposta em processos críticos. Essa integração de raciocínio automatizado e interfaces conversacionais robustas evidencia o potencial dos LLMs para transformar setores intensivos em conhecimento, ao mesmo tempo em que impõe novos desafios de robustez, sensibilidade a erros de entrada e necessidade de integração com sistemas legados, aspectos detalhados por Cremer (2025) em seus estudos de caso.

Esta realidade revela o que denominamos de "paradoxo da cognição artificial escalonável": um dilema fundamental onde o aumento das capacidades cognitivas dos LLMs através de maior escala de parâmetros intensifica exponencialmente as demandas por recursos computacionais, criando uma tensão irreconciliável entre capacidade emergente e sustentabilidade computacional. Este paradoxo estabelece um trade-off crítico que força uma reconsideração das métricas de sucesso em IA: devemos priorizar capacidades absolutas ou eficiência relativa? A resposta a esta questão tem implicações diretas para a democratização do acesso a tecnologias cognitivas avançadas e constitui o núcleo da motivação desta pesquisa. Em sintonia com esse dilema, Jiang e Xia (2025) discutem como a integração de LLMs em navegadores inteligentes, como o Orca, pode mitigar parte dessa tensão ao promover uma orquestração colaborativa entre usuário e IA. Ao invés de substituir o humano por automação total, o Orca explora interfaces maleáveis e automação sob demanda para reduzir o esforço cognitivo e manual, preservando a agência do usuário e otimizando o uso dos recursos computacionais. Assim, a abordagem defendida por Jiang e Xia sugere que o futuro da IA escalável pode residir em sistemas híbridos, onde a eficiência emerge da colaboração dinâmica entre capacidades humanas e assistentes inteligentes, ao invés de depender exclusivamente do aumento indiscriminado de escala computacional.

A evolução paradigmática dos LLMs, conforme observada na plataforma CrossDebate, demonstra empiricamente a necessidade de convergência disciplinar através dos dados coletados em nosso estudo. Os resultados específicos do monitoramento EEG revelaram que a potência da banda Téta frontal aumentou significativamente durante operações com modelos Q4, F(1, 38) = 12.40, p < .001, η²p = .25, 95% CI [.08, .42], representando aumento médio de 15.3 μV²/Hz, 95% CI [12.1, 18.5] em Q4 versus 8.7 μV²/Hz, 95% CI [6.5, 10.9] em Q8, corroborando princípios da neurociência cognitiva sobre processamento de memória de trabalho sob sobrecarga. As métricas de sistema registradas via psutil mostraram que os frameworks de otimização de recursos da engenharia de software foram eficazes em reduzir uso de VRAM em 40-50% com quantização Q4, validando abordagens de sistemas distribuídos.

Particularmente relevante, nossos dados de NASA-TLX demonstraram aplicabilidade direta da Teoria da Carga Cognitiva de Sweller (2011), com escores significativamente elevados (η²p ≈ 0.40) para condições Q4, especificamente nas subescalas de Demanda Mental e Frustração. Os protocolos de pensamento em voz alta revelaram que 87% dos participantes relataram necessidade de "verificação constante" com Q4, fornecendo evidência comportamental específica para sobrecarga da memória de trabalho postulada pela CLT.

Nossos dados sugerem que esta convergência disciplinar pode ser essencial para o desenvolvimento de sistemas de AI locais mais ergonômicos, embora sejam necessários estudos longitudinais para validar a generalização destes achados além do contexto específico da plataforma CrossDebate com modelos GGUF.

As sessões experimentais de 60 minutos foram estrategicamente delineadas seguindo princípios de design experimental balanceado para capturar com alta fidelidade as respostas psicofisiológicas agudas às manipulações experimentais específicas (Q4 vs Q8, complexidade multiagente). Esta duração representa um compromisso metodológico otimizado entre três critérios fundamentais: (a) tempo suficiente para exposição confiável a todas as condições experimentais com pelo menos 15 minutos por condição; (b) manutenção da qualidade dos sinais psicofisiológicos sem degradação por fadiga ou artefatos instrumentais; e (c) controle rigoroso de variáveis confundidoras relacionadas à adaptação estratégica de longo prazo que poderia obscurecer os efeitos específicos da interdependência CL-CompL investigada. Esta duração representa um ponto ótimo metodológico que equilibra profundidade de medição com controle experimental, priorizando a caracterização precisa da dinâmica inicial CL-CompL sobre observações de adaptação estratégica de longo prazo. Embora esta escolha temporal delimite nosso escopo às respostas táticas-operacionais imediatas, ela estabelece fundamentos empíricos essenciais para estudos longitudinais futuros sobre sustentabilidade da interação humano-IA em contextos profissionais prolongados, onde processos adaptativos de escalas temporais distintas podem modular diferentemente a interdependência cognitivo-computacional.

O fenômeno das habilidades emergentes em LLMs levanta questões fundamentais sobre a natureza da inteligência artificial e suas implicações para a interação humano-computador. Quando modelos como Claude 4 ou LLaMA-3 demonstram capacidades de raciocínio que não estavam presentes em seus predecessores menores, observamos não apenas um salto quantitativo, mas uma transformação qualitativa na natureza da computação. Esta emergência, contudo, traz consigo novos desafios: como podemos prever, controlar e otimizar sistemas cujas capacidades emergem de forma não linear? Como garantimos que essas capacidades emergentes sejam alinhadas com objetivos humanos e utilizáveis de forma ergonômica? Estas questões tornam-se ainda mais complexas quando consideramos que a emergência não ocorre apenas no nível do modelo individual, mas também no nível de sistemas multiagente, onde a orquestração de múltiplos LLMs pode produzir comportamentos coletivos inesperados.

Historicamente, no entanto, a própria escala que confere poder a esses LLMs representou uma importante barreira. Os vastos recursos computacionais (medidos em PetaFLOP-dias), de memória (centenas de gigabytes VRAM/RAM) e energéticos necessários para treinar e, em menor grau, para servir modelos de fronteira, confinaram seu desenvolvimento e uso a um número limitado de grandes corporações tecnológicas e instituições de pesquisa com recursos substanciais (Strubell et al., 2019; Bender et al., 2021; Bai et al., 2024). Essa centralização levantou preocupações sobre equidade no acesso, reprodutibilidade da pesquisa e a pegada ambiental da AI (Husom et al., 2025), limitando a exploração e inovação por parte de pesquisadores independentes, pequenas empresas e estudantes.

A centralização dos LLMs em grandes corporações criou um paradoxo na democratização da inteligência artificial. Por um lado, essas organizações possuem os recursos necessários para desenvolver modelos cada vez mais poderosos; por outro, essa concentração de poder computacional estabelece barreiras significativas para inovação distribuída e experimentação independente. Este cenário é reminiscente dos primórdios da computação, quando mainframes eram acessíveis apenas a grandes instituições, até que o advento dos computadores pessoais revolucionou o campo. A atual transição para LLMs locais representa uma mudança paradigmática similar, com implicações profundas não apenas para a tecnologia, mas para a economia, educação e sociedade como um todo.

As preocupações ambientais associadas aos LLMs de grande escala estendem-se além do consumo energético direto para incluir considerações sobre sustentabilidade a longo prazo da pesquisa em AI. O treinamento de um único modelo de fronteira pode consumir a energia equivalente ao uso doméstico de centenas de famílias por um ano inteiro, levantando questões éticas sobre o custo ambiental da busca por inteligência artificial avançada. Esta realidade torna ainda mais urgente o desenvolvimento de alternativas eficientes, como os modelos quantizados GGUF, que podem reduzir drasticamente o consumo energético sem comprometer significativamente a funcionalidade. A democratização através de hardware local não é, portanto, apenas uma questão de acessibilidade, mas também de responsabilidade ambiental e sustentabilidade.

Os avanços revolucionários em técnicas de quantização, particularmente a transição de representações INT8 para INT4, têm demonstrado capacidade de reduzir drasticamente o uso de VRAM em até 75%, transformando fundamentalmente a viabilidade de execução local de Grandes Modelos de Linguagem. Esta redução significa que modelos com 32 bilhões de parâmetros, anteriormente executáveis apenas em infraestruturas corporativas com centenas de gigabytes de memória dedicada, podem agora operar eficientemente em GPUs modestas de consumidor, como RTX 3060 ou 4050, com apenas 6-12GB de VRAM. A padronização de formatos otimizados como GGUF (GPT-Generated Unified Format), combinada com runtimes especializados como llama.cpp, criou um ecossistema robusto que não apenas viabiliza, mas otimiza a execução local em hardware acessível. Esta convergência tecnológica está acelerando exponencialmente a democratização dos LLMs, removendo barreiras econômicas e técnicas que historicamente limitavam o acesso à inteligência artificial avançada a grandes corporações e instituições de pesquisa bem financiadas. A quantização representa muito mais do que uma simples otimização técnica; ela constitui uma mudança fundamental na filosofia de desenvolvimento de AI. Tradicionalmente, o paradigma dominante enfatizava a maximização da precisão através do aumento de parâmetros e dados, independentemente do custo computacional. A quantização inverte essa lógica, priorizando a eficiência e acessibilidade enquanto mantém funcionalidade aceitável. Esta mudança de paradigma é particularmente significativa quando consideramos que a precisão perfeita nem sempre é necessária para aplicações práticas - frequentemente, uma redução modesta na precisão é um balanço aceitável pelos ganhos substanciais em acessibilidade e eficiência. O formato GGUF exemplifica essa filosofia, otimizando não apenas para redução de memória, mas também para carregamento rápido e compatibilidade multiplataforma.

O conceito de "on-device AI" introduz uma nova dimensão na arquitetura de sistemas inteligentes, movendo-se de um modelo centralizado baseado em nuvem para um modelo distribuído onde a inteligência reside localmente nos dispositivos dos usuários. Esta transição tem implicações profundas para privacidade, latência, personalização e autonomia do usuário. Quando um LLM roda localmente, os dados do usuário nunca precisam deixar o dispositivo, eliminando preocupações sobre privacidade e permitindo aplicações verdadeiramente personalizadas. Além disso, a ausência de dependência de conectividade de rede torna essas aplicações mais robustas e confiáveis, especialmente em contextos onde a conectividade é limitada ou instável. Essa descentralização tecnológica dialoga diretamente com as discussões de Constantinides e Quercia (2025) sobre a necessidade de sistemas de IA que priorizem a agência humana e a personalização, em oposição ao paradigma dominante de automação centralizada. Os autores argumentam que abordagens centradas no usuário, como a IA local, podem promover maior alinhamento com valores sociais e cognitivos, além de facilitar a participação ativa dos usuários no controle e adaptação dos sistemas. Assim, a adoção de "on-device AI" não apenas reforça a privacidade, mas também potencializa a construção de sistemas de IA verdadeiramente augmentativos, que ampliam capacidades humanas ao invés de simplesmente automatizar tarefas, conforme defendido no framework de Human-Centered AI (HCAI) proposto por Constantinides e Quercia (2025).

A quantização representa uma das técnicas mais fundamentais e transformadoras na otimização de redes neurais profundas, operando através da redução sistemática e controlada da precisão numérica dos pesos do modelo. O processo transforma representações de ponto flutuante de alta precisão (tipicamente FP32 com 32 bits ou FP16 com 16 bits) para representações de inteiros de menor precisão (como INT8 com 8 bits, INT4 com 4 bits, ou até mesmo representações binárias de 1 bit). Este processo de compressão numérica diminui substancialmente o uso de memória VRAM necessário para armazenar os parâmetros do modelo, frequentemente resultando em reduções de 50-75% nos requisitos de memória dependendo do nível de quantização aplicado. Simultaneamente, a quantização acelera significativamente a inferência ao permitir que operações matemáticas sejam executadas em unidades aritméticas de menor precisão, que são inerentemente mais rápidas e energeticamente eficientes.

A universalidade deste desafio de otimização é evidenciada em domínios aparentemente distantes, como demonstrado por Ruzzelli et al. (2010) em sistemas de desagregação de energia, onde o trade-off fundamental entre eficiência computacional e acurácia se manifesta de forma análoga ao dilema Q4/Q8 em LLMs. Em tais sistemas, o aumento de variáveis pode sobrecarregar os algoritmos, impactando a interatividade e degradando a experiência do usuário, um fenômeno que ecoa diretamente os desafios enfrentados na quantização agressiva de modelos neurais, onde a redução excessiva da precisão pode comprometer a qualidade das saídas e aumentar a carga cognitiva do operador.

A estratégia de "sleep-time compute", introduzida por Lin et al. (2025), oferece uma abordagem complementar inovadora à quantização tradicional. Enquanto a quantização otimiza o modelo através da redução de precisão, o sleep-time compute otimiza o processamento temporal, permitindo que modelos pré-processem contexto offline para reduzir dramaticamente a carga computacional durante a inferência em tempo real. Esta técnica se torna particularmente relevante para sistemas como a plataforma CrossDebate, onde agentes GGUF poderiam realizar "cálculos em tempo de sono" sobre documentos a serem resumidos ou tópicos de debate, amortizando custos computacionais através de múltiplas consultas relacionadas, como demonstrado no dataset Multi-Query GSM-Symbolic.

Contudo, esta otimização introduz um balanço fundamental: enquanto os benefícios computacionais são substanciais, a redução na precisão numérica pode resultar em degradação da qualidade das saídas do modelo, criando uma tensão entre eficiência computacional e fidelidade das respostas que deve ser cuidadosamente balanceada durante o design de sistemas de AI práticos. Este equilíbrio ecoa o conceito de racionalidade limitada proposto por Herbert Simon (1957), onde tanto sistemas humanos quanto artificiais não otimizam globalmente, mas "satisfazem" - encontrando soluções que são "boas o suficiente" para as demandas contextuais específicas.

Formatos como GGUF, aliados a runtimes otimizados como llama.cpp e frameworks como Ollama, permitem executar LLMs bilionários em hardware comum, como laptops com GPUs RTX 3060 ou 4050. Essa execução local ("on-device AI") reduz custos e latência, viabilizando aplicações de AI personalizadas e offline em dispositivos acessíveis. A democratização proporcionada por esses formatos otimizados representa uma ruptura fundamental no paradigma tradicional de acesso à AI avançada, onde apenas organizações com recursos substanciais podiam deployar modelos de grande escala. A combinação de quantização eficiente, otimizações de runtime específicas para hardware consumer e protocolos padronizados de armazenamento não apenas torna viável a execução local, mas também estabelece um ecossistema sustentável onde desenvolvedores independentes, pesquisadores acadêmicos e pequenas empresas podem experimentar, iterar e inovar com tecnologias que anteriormente estavam restritas a datacenters corporativos. Esta transformação tem implicações profundas para privacidade de dados, soberania digital e inovação distribuída, criando novas oportunidades para aplicações especializadas que requerem processamento local e personalização fina.

Embora a capacidade de executar inferência com modelos GGUF pré-treinados localmente represente um avanço relevante, a realização de seu pleno potencial para aplicações especializadas frequentemente exige ajuste fino (*fine-tuning*) (Howard e Ruder, 2018; Raffel et al., 2020). Usando um conjunto de dados menor e direcionado, o ajuste fino adapta um modelo pré-treinado a uma tarefa, por exemplo, geração de código para um framework em particular (Afrin et al., 2025; Ayupov e Chirkova, 2022). Ele permite que os usuários personalizem o comportamento do LLM, superando as limitações do aprendizado *in-context* para tarefas que exigem aquisição de conhecimento de nicho (Jindal et al., 2024; Saroufim et al., 2025).

O ajuste fino representa a diferença entre ter acesso a uma ferramenta genérica e possuir um instrumento especializado perfeitamente calibrado para necessidades específicas. Enquanto modelos pré-treinados demonstram capacidades impressionantes em tarefas gerais, eles frequentemente carecem da especificidade e precisão necessárias para domínios altamente especializados. Por exemplo, um LLM genérico pode ter conhecimento básico sobre medicina, mas um modelo ajustado finamente com literatura médica especializada pode oferecer insights clínicos muito mais precisos e úteis. Esta especialização através do ajuste fino é particularmente valiosa em contextos educacionais, onde diferentes disciplinas requerem não apenas conhecimento factual, mas também metodologias de raciocínio e estilos de comunicação específicos.

A limitação do aprendizado in-context, embora significativa, também revela aspectos fundamentais sobre como os LLMs processam e utilizam informações. O aprendizado in-context permite que modelos adaptem seu comportamento com base em exemplos fornecidos no prompt, mas essa adaptação é limitada pelo contexto da janela de entrada e pela capacidade do modelo de generalizar a partir de poucos exemplos. O ajuste fino, por outro lado, permite uma adaptação mais profunda e permanente, modificando os próprios pesos do modelo para incorporar novo conhecimento. Esta distinção é primordial para compreender quando e como usar cada abordagem, especialmente em contextos onde a consistência e confiabilidade são prioritárias.

No entanto, o ajuste fino completo (FFT), que atualiza todos os parâmetros do modelo, permanece computacionalmente proibitivo para LLMs robustos em hardware local, mesmo quando o modelo base é quantizado, devido aos requisitos de memória para armazenar gradientes e ativações intermediárias para retropropagação (Dodge et al., 2022; Afrin et al., 2025).

As limitações computacionais do ajuste fino completo não são meramente técnicas, mas fundamentalmente matemáticas. Durante o treinamento, cada parâmetro do modelo requer não apenas espaço para seu valor atual, mas também para seu gradiente, momento, e outras informações de otimização. Para um modelo com bilhões de parâmetros, isso pode resultar em requisitos de memória que excedem por ordens de magnitude a capacidade de hardware consumer. Além disso, o processo de retropropagação requer o armazenamento de ativações intermediárias para cada camada, criando uma demanda de memória que cresce linearmente com o tamanho do modelo e o tamanho do batch. Esta realidade matemática torna o ajuste fino completo impraticável não apenas por limitações de hardware, mas por limitações fundamentais de escalabilidade.

Essa limitação impulsionou o desenvolvimento e a adoção generalizada de técnicas de Ajuste Fino Eficiente em Parâmetros (PEFT) (Houlsby et al., 2019; Lester et al., 2021; Lialin et al., 2023; Afrin et al., 2025; Wang et al., 2025b; Weyssow et al., 2023). Métodos PEFT, como a proeminente Low-Rank Adaptation (LoRA) (Hu et al., 2022) e suas variantes, adaptam LLMs pré-treinados adicionando novos parâmetros treináveis de baixa dimensionalidade, mantendo a vasta maioria dos pesos originais congelados. A Adaptação de Baixo Posto Quantizada (QLoRA) (Dettmers et al., 2023; Afrin et al., 2025), em particular, combina LoRA com quantização agressiva (4 bits, como o formato INT4) do modelo base congelado e otimizações de memória adicionais (Double Quantization). Essa sinergia tornou possível ajustar finamente modelos GGUF de dezenas de bilhões de parâmetros em GPUs acessíveis com VRAM limitada (por exemplo, 6-12GB, como as RTX 3060 e 4050), representando um marco na democratização não apenas da inferência, mas também da adaptação de LLMs localmente. Bibliotecas como peft 0.11.0 da HuggingFace (Saroufim et al, 2025) e otimizações como unsloth 2024.8 facilitaram a implementação prática dessas técnicas, tornando o ajuste fino local uma realidade tangível para um público mais amplo.

As técnicas PEFT representam uma elegante solução ao dilema da especificidade versus eficiência. A genialidade da abordagem LoRA reside na observação matemática de que as mudanças necessárias para especializar um modelo pré-treinado podem frequentemente ser representadas como matrizes de baixo posto. Esta insight não é apenas praticamente útil, mas teoricamente profundo, sugerindo que a adaptação de domínio pode ser capturada por um subespaço relativamente pequeno do espaço total de parâmetros. Esta descoberta tem implicações que vão além da eficiência computacional, oferecendo insights sobre a natureza da transferência de aprendizado e a estrutura do conhecimento em redes neurais.

A combinação sinérgica de LoRA com quantização em QLoRA exemplifica como inovações aparentemente independentes podem convergir para criar capacidades transformadoras. A quantização reduz os requisitos de memória do modelo base, enquanto LoRA minimiza os parâmetros adicionais necessários para adaptação. Juntas, essas técnicas tornam possível o que antes era impensável: ajustar finamente modelos de dezenas de bilhões de parâmetros em hardware consumer. Esta convergência não é acidental, mas representa uma tendência mais ampla na pesquisa de AI em direção a soluções que priorizam eficiência e acessibilidade sem comprometer funcionalidade.

O desenvolvimento de bibliotecas especializadas como peft e otimizações como unsloth representa a maturação do ecossistema de ajuste fino eficiente. Estas ferramentas não apenas implementam algoritmos complexos, mas os tornam acessíveis a uma audiência mais ampla através de APIs intuitivas e documentação abrangente. A disponibilidade dessas ferramentas marca a transição do ajuste fino eficiente de uma técnica experimental para uma prática padrão, democratizando não apenas o acesso aos modelos, mas também a capacidade de adaptá-los para necessidades específicas.

O ajuste fino local de LLMs GGUF exige supervisão constante do operador, que precisa curar dados, criar prompts, ajustar hiperparâmetros e monitorar o desempenho. Sob uma Articulação Teórica Transdisciplinar, propomos integrar a Cognitive Load Theory (Sweller, 2011), que postula sobre os limites da memória de trabalho humana durante tarefas exigentes, com frameworks de avaliação ergonômica como a ISO 9241-210:2019, que estabelece princípios de design centrado no ser humano para sistemas interativos. Essa articulação permite analisar a complexidade do ajuste fino local considerando tanto as limitações cognitivas do operador (CL) quanto os requisitos ergonômicos e de usabilidade da interface e dos processos, promovendo uma interação humano-computador mais eficaz, eficiente e satisfatória, mesmo diante da complexidade inerente à tarefa.

A necessidade de supervisão humana intensiva no ajuste fino local revela uma tensão fundamental entre autonomia tecnológica e controle humano. Embora sistemas automatizados possam executar muitas tarefas de forma independente, o ajuste fino requer julgamentos qualitativos sobre dados, parâmetros e resultados que dependem de expertise humana e compreensão contextual. Esta realidade destaca a importância de projetar sistemas que não substituam a expertise humana, mas a amplificam e a tornam mais eficiente. A Articulação Teórica Transdisciplinar proposta reconhece que esta colaboração humano-máquina deve ser fundamentada em princípios tanto cognitivos quanto ergonômicos para ser verdadeiramente eficaz.

A integração da Cognitive Load Theory com frameworks ergonômicos representa uma abordagem inovadora para design de sistemas de AI. A CLT fornece insights sobre como os seres humanos processam informações complexas, identificando três tipos de carga cognitiva: intrínseca (relacionada à complexidade inerente da tarefa), extrínseca (relacionada à forma como a informação é apresentada) e germânica (relacionada ao processo de construção de esquemas mentais). Quando aplicada ao design de interfaces para ajuste fino de LLMs, esta teoria sugere que devemos minimizar a carga extrínseca através de design intuitivo, enquanto otimizamos a carga germânica para facilitar o aprendizado e a expertise. Estudos como o de Budal e Ebbesen (2024), que investigaram sistemas automatizados de baixo custo sob uma perspectiva centrada no humano, demonstraram como a sobrecarga computacional pode degradar significativamente a experiência do operador, corroborando a necessidade crítica de balancear eficiência técnica com usabilidade. A ISO 9241-210:2019 complementa esta perspectiva oferecendo diretrizes práticas para implementar princípios centrados no usuário.

O ajuste fino de modelos GGUF quantizados com QLoRA impõe uma vultosa Carga Computacional (CompL) ao hardware local, exigindo gerenciamento eficiente de VRAM, processamento e energia. Esta demanda técnica, conforme a Articulação Teórica Transdisciplinar que propomos, relaciona-se diretamente com a Carga Cognitiva (CL) do operador (Sweller, 2011), que tem de equilibrar eficiência e precisão ao gerenciar sistemas em hardware com restrições. A norma ISO 9241-210:2019, que preconiza o design centrado no ser humano, é essencial aqui, pois uma CompL excessiva ou mal gerenciada pode levar a falhas de tempo de execução ou lentidão da plataforma, como observado em outras aplicações com alta CompL (Budal e Ebbesen, 2024), impactando negativamente a usabilidade. Tal cenário aumenta a CL extrínseca, pois o operador necessita despender esforço adicional para contornar problemas ou esperar pela plataforma, distanciando-se dos objetivos de eficácia, eficiência e satisfação preconizados pela ISO 9241-210:2019. A otimização da CompL, como a não adição de bibliotecas de física desnecessárias (Rodriguez-Ramirez et al., 2025) ou a redução da dimensionalidade de características (conforme Kango et al., 2024), torna-se um requisito não apenas técnico, mas ergonômico, para alinhar a interação com os princípios da ISO 9241-210:2019, minimizando a CL imposta pela máquina.

A interdependência entre Carga Computacional e Carga Cognitiva revela-se particularmente complexa em sistemas de hardware com recursos limitados. Quando um sistema opera próximo aos limites de sua capacidade computacional, pequenas flutuações na demanda podem resultar em degradação perceptível de desempenho, forçando o operador a adaptar constantemente suas estratégias de trabalho. Esta adaptação contínua consome recursos cognitivos preciosos que poderiam ser direcionados para a tarefa primária. O resultado é um ciclo de feedback negativo onde limitações computacionais amplificam a carga cognitiva, que por sua vez pode levar a decisões sub-ótimas que aumentam ainda mais a demanda computacional.

A aplicação dos princípios da ISO 9241-210:2019 ao contexto de sistemas de AI locais requer uma reinterpretação cuidadosa dos conceitos tradicionais de usabilidade. Eficácia, eficiência e satisfação assumem novas dimensões quando consideramos sistemas que não apenas respondem a comandos do usuário, mas que também aprendem e se adaptam ao longo do tempo. A eficácia deve ser medida não apenas pela capacidade de completar tarefas imediatas, mas pela qualidade do aprendizado e adaptação do sistema. A eficiência deve considerar não apenas o tempo necessário para completar tarefas, mas também o investimento cognitivo requerido do usuário. A satisfação deve englobar não apenas a experiência imediata, mas a confiança a longo prazo no sistema e sua evolução.

A natureza não-linear da relação CL-CompL torna-se evidente quando consideramos cenários de otimização extrema. Em muitos casos, tentativas de minimizar drasticamente a CompL através de quantização agressiva ou redução de parâmetros podem resultar em aumentos desproporcionais na CL, conforme o operador luta para compensar a redução na qualidade do modelo. Conversely, investimentos computacionais que excedem as capacidades do hardware disponível podem criar gargalos que frustram o usuário e aumentam a carga cognitiva através de delays e instabilidade. O desafio reside em identificar o ponto de equilíbrio onde a redução na CompL não compromete significativamente a qualidade da saída, e onde os recursos computacionais são utilizados de forma eficiente sem sobrecarregar o sistema.

A complexidade exponencial introduzida por sistemas multiagente adiciona uma dimensão qualitativemente diferente ao problema CL-CompL. Quando múltiplos agentes operam simultaneamente, emergem fenômenos de coordenação, competição por recursos e sincronização que não existem em sistemas de agente único. A orquestração de 60 agentes GGUF, como implementada neste estudo, requer não apenas o gerenciamento de recursos computacionais para cada agente individual, mas também a coordenação inteligente de suas interações. Esta orquestração demanda do operador uma compreensão sistêmica que vai além do conhecimento técnico individual de cada componente, exigindo capacidades de pensamento sistêmico e gestão de complexidade que representam uma forma especializada de carga cognitiva.

A plataforma CrossDebate orquestrou 60 modelos GGUF (30 Q4, 30 Q8) com ajuste fino QLoRA e monitoramento em tempo real. Suas aplicações incluíram resumo colaborativo de documentos e simulações educacionais de debates. Diferentemente de Dettmers et al. (2023), nossa pesquisa correlacionou QLoRA com métricas fisiológicas de Carga Cognitiva (CL), como EEG e pupilometria, quantificando o impacto humano da quantização. Esta abordagem, fundamentada na Articulação Teórica Transdisciplinar, integrou a Cognitive Load Theory (Sweller, 2011) com métricas de desempenho computacional (CompL). A plataforma, com interface React (pacotes react 18.3.1, chart.js 4.4.9, express 4.18.2, puppeteer 24.8.0, react-dom 18.3.1) e backend FastAPI (pacotes fastapi 0.115.6 e uvicorn 0.21.1), integrou dados psicofisiológicos e computacionais para otimizar a interação CL-CompL, conseguindo criar plataformas de AI locais mais ergonômicas, em consonância com os princípios da ISO 9241-210:2019, que enfatiza a usabilidade e a experiência do usuário como centrais.

A escolha de 60 agentes distribuídos entre configurações Q4 e Q8 não foi arbitrária, mas baseada em considerações teóricas sobre diversidade e redundância em sistemas complexos. A teoria de sistemas sugere que a robustez emerge da diversidade de componentes e de suas interações, enquanto a redundância fornece tolerância a falhas. Os 30 agentes Q4 oferecem velocidade e eficiência computacional, enquanto os 30 agentes Q8 fornecem precisão e qualidade superior. Esta distribuição permite que o sistema adapte dinamicamente sua estratégia, utilizando agentes Q4 para tarefas que requerem rapidez e agentes Q8 para tarefas que demandam precisão, criando uma arquitetura híbrida que otimiza tanto para performance quanto para qualidade.

A integração de métricas fisiológicas representa uma inovação significativa na avaliação de sistemas de AI. Tradicionalmente, a avaliação de interfaces e sistemas focava em métricas objetivas de performance ou avaliações subjetivas através de questionários. A incorporação de EEG e pupilometria adiciona uma dimensão biológica que pode revelar aspectos da experiência do usuário que não são conscientes ou facilmente verbalizáveis. Por exemplo, a pupilometria pode detectar carga cognitiva mesmo quando o usuário não está ciente de estar experimentando dificuldade, enquanto o EEG pode revelar padrões de atenção e processamento que precedem problemas de performance. Esta abordagem multi-modal oferece uma visão mais completa e objetiva da interação humano-computador.

A arquitetura tecnológica da plataforma CrossDebate reflete princípios modernos de desenvolvimento de software que priorizam modularidade, escalabilidade e manutenibilidade. A escolha do React para o frontend permite interfaces dinâmicas e responsivas que podem adaptar-se em tempo real às mudanças no estado do sistema, enquanto o FastAPI no backend oferece performance excepcional e documentação automática de APIs. Esta separação clara entre frontend e backend não apenas facilita o desenvolvimento e manutenção, mas também permite que diferentes aspectos do sistema sejam otimizados independentemente, um princípio essencial quando se trabalha com sistemas que devem balancear múltiplas demandas de performance.

A CrossDebate integra dados psicofisiológicos (EEG, POG, PPG) e computacionais (uso de VRAM, energia) para monitorar simultaneamente a Carga Cognitiva do usuário (CL) e a Carga Computacional (CompL) durante operações de ajuste fino com QLoRA. Através de uma interface React e backend FastAPI, a plataforma monitora em tempo real o desempenho dos modelos e o estado do operador, possibilitando ajustes dinâmicos nos parâmetros e fluxos de trabalho multiagente. A hipótese central é que visualizar e otimizar a relação CL-CompL permite criar sistemas de AI locais mais ergonômicos e eficientes para hardware de prateleira.

O monitoramento simultâneo de métricas psicofisiológicas e computacionais cria oportunidades para sistemas adaptativos que respondem automaticamente ao estado do usuário. Quando o sistema detecta indicadores de alta carga cognitiva através do EEG ou pupilometria, pode automaticamente ajustar parâmetros como complexidade de visualização, frequência de atualizações ou nível de detalhe nas informações apresentadas. Similarmente, quando métricas computacionais indicam stress no sistema, a interface pode proativamente sugerir otimizações ou automaticamente implementar estratégias de redução de carga. Esta capacidade de adaptação bidirecional representa uma evolução significativa em direção a sistemas verdadeiramente centrados no usuário.

A hipótese central da pesquisa fundamenta-se na premissa de que a visualização da relação CL-CompL pode transformar a natureza da interação humano-computador em sistemas de AI. Tradicionalmente, usuários operam com visibilidade limitada sobre o estado interno de sistemas complexos, forçando-os a fazer inferências baseadas em indicadores indiretos de performance. Ao tornar visível a relação entre carga cognitiva e computacional, a plataforma permite que usuários tomem decisões mais informadas sobre como balancear eficiência e qualidade, desenvolver estratégias de trabalho mais eficazes e identificar oportunidades de otimização que não seriam aparentes de outra forma.

A democratização da inteligência artificial através de hardware local representa não apenas uma mudança tecnológica, mas uma transformação social que pode redefinir quem tem acesso às ferramentas mais poderosas da nossa era. Quando LLMs avançados podem rodar em laptops comuns, rompem-se barreiras geográficas, econômicas e institucionais que tradicionalmente limitavam o acesso à AI de ponta. Esta democratização tem potencial para acelerar inovação em regiões e comunidades anteriormente excluídas do desenvolvimento de AI, criando um ecossistema mais diverso e inclusivo de pesquisa e aplicação.

As implicações dessa democratização estendem-se para além da tecnologia para questões fundamentais de autonomia, privacidade e soberania digital. Quando indivíduos e organizações podem operar sistemas de AI avançados localmente, reduzem sua dependência de serviços centralizados e grandes corporações tecnológicas. Esta independência tecnológica é particularmente importante para aplicações sensíveis, pesquisa confidencial e contextos onde a conectividade de rede é limitada ou insegura.

A capacidade de ajustar finamente modelos localmente adiciona uma dimensão adicional de personalização e controle que é impossível com sistemas baseados em nuvem. Este controle local manifesta-se através de um padrão em U na relação entre Carga Computacional (CompL) e Carga Cognitiva (CL): configurações de recursos computacionais muito baixas (hardware insuficiente, modelos excessivamente quantizados) resultam em alta CL devido à degradação da qualidade e instabilidade do sistema; configurações excessivamente altas (hardware sobredimensionado, modelos não-otimizados) também elevam a CL através de sobrecarga de opções e complexidade desnecessária; o ponto ótimo intermediário minimiza a CL ao equilibrar capacidade computacional adequada com simplicidade operacional, permitindo personalização eficiente sem sacrificar usabilidade.

Nosso objetivo é demonstrar que a análise detalhada dos processos de raciocínio, como os implementados em modelos como DeepSeek-R1 (Marjanovic et al., 2025), pode oferecer insights valiosos sobre as fontes de Carga Cognitiva (CL) (Sweller, 2011), como raciocínio ineficiente, e sua correlação com a Carga Computacional (CompL), utilizando ferramentas especializadas na plataforma CrossDebate. O formato GGUF, um padrão otimizado para modelos quantizados, garante carregamento rápido e compatibilidade com frameworks como llama.cpp, que executa LLMs em CPUs, reduzindo a dependência de GPUs. O FastAPI, um framework Python, suporta a orquestração de 60 agentes e registrar métricas em tempo real, enquanto o React criou interfaces dinâmicas para visualizar CL e CompL.

Esta visualização e a capacidade de registrar métricas foram fundamentais sob a perspectiva da ISO 9241-210:2019, que advoga por um design iterativo e centrado no usuário, onde a compreensão da experiência do usuário (incluindo CL) e da eficiência da plataforma (influenciada pela CompL) informou melhorias. A relação entre CompL e CL manifesta-se através de um padrão em J característico: configurações de recursos computacionais baixas a moderadas mantêm a CL em níveis relativamente estáveis devido à capacidade adaptativa dos operadores; contudo, quando a CompL atinge limiares críticos - como no caso de algoritmos que se tornam sobrecarregados com o aumento de variáveis (Ruzzelli et al., 2010) ou quantização excessivamente agressiva que degrada a qualidade dos outputs - observa-se uma escalada exponencial irreversível da CL que degrada drasticamente a interatividade e prejudica a usabilidade sem possibilidade de retorno a estados eficientes anteriores. Este padrão J caracteriza-se pela ausência de um ponto ótimo de recuperação, onde a CL permanece gerenciável até que gargalos computacionais críticos disparem mecanismos de sobrecarga cognitiva que forçam estratégias compensatórias custosas, amplificando ainda mais a demanda mental. Conseguimos quantificar como decisões de design, como a quantização GGUF ou a complexidade de fluxos multiagente, impactaram esta dinâmica não-linear, identificando configurações que evitaram os limiares críticos do padrão J, conforme preconizado pela ISO 9241-210:2019, para promover interações ergonômicas que reconhecem os limites adaptativos da cognição humana em sistemas de AI locais.

Nossa comparação entre os níveis de quantização mostrou que os modelos Q4 oferecem maior eficiência computacional (37.5 ± 2.5 tokens/s em RTX 3060, metade do consumo de VRAM) em relação aos Q8 (17.5 ± 1.25 tokens/s), porém com qualidade inferior (degradação de 2-4% versus <1%). A implementação de fluxos multiagente na plataforma CrossDebate reduziu o tempo médio de resposta em 18%, demonstrando o potencial de plataformas GGUF locais para tarefas complexas. A plataforma orquestrou eficientemente 60 modelos (30 Q4, 30 Q8) em hardware acessível, com integração de dados psicofisiológicos e computacionais para equilibrar Carga Cognitiva (CL) e Computacional (CompL), proporcionando maior usabilidade em aplicações educacionais e analíticas.

Na educação superior, agentes LLM Q8 corrigem redações com alta precisão, enquanto modelos Q4 em dispositivos móveis necessitam revisão humana em 15% dos casos devido a ambiguidades. Embora GGUF e sistemas multiagente sejam cada vez mais adotados, a relação entre Carga Cognitiva (CL) e Computacional (CompL) permanece pouco explorada. Nossa pesquisa quantifica como a orquestração de 60 agentes (30 Q4/Q8) afeta a eficiência e ergonomia em aplicações educacionais.

Pesquisas anteriores frequentemente otimizam a Carga Computacional (CompL) ou estudam a Carga Cognitiva (CL) isoladamente. Nosso trabalho, ao adotar uma Articulação Teórica Transdisciplinar, busca superar essa limitação integrando modelos como a Cognitive Load Theory (Sweller, 2011) com métricas de desempenho computacional e frameworks de avaliação ergonômica como a ISO 9241-210:2019. Quantificamos esse custo humano (via EEG/questionários), revelando que modelos Q4 aumentam a CL pela menor precisão. Integrando dados psicofisiológicos e computacionais na CrossDebate, analisamos a interdependência CL-CompL em 60 agentes GGUF (Q4/Q8) em hardware comum. Esta abordagem integrada é essencial, pois a ISO 9241-210:2019 enfatiza a necessidade de um design que considere toda a experiência do usuário, o que inerentemente envolve o impacto da CompL na CL e, consequentemente, na eficácia, eficiência e satisfação ao usar sistemas de AI locais ergonômicos.

Nunca antes um estudo sobre ajuste fino HITL vinculou diretamente às compensações de CompL (Amershi et al., 2019; Stiennon et al., 2020). As pesquisas em sistemas multiagente negligenciam o fardo combinado CL/CompL sobre o orquestrador humano em ambientes com recursos limitados (Wu et al., 2023; Hong et al., 2023; Luo et al., 2025a), focando mais na coordenação entre agentes do que na interação com o operador humano (OpenAI, 2025). Além disso, estratégias emergentes para melhorar modelos menores usando saídas de modelos maiores e mais capazes (como os modelos de raciocínio) (Wang et al., 2025a) adicionam outra dimensão a ser considerada: como o processo de destilação de conhecimento ou ajuste fino supervisionado com dados de alta qualidade afeta o equilíbrio CL-CompL?

A democratização dos Grandes Modelos de Linguagem (LLMs), impulsionada por formatos otimizados como GGUF e técnicas eficientes como QLoRA, justifica a relevância deste estudo, que visa preencher a lacuna crítica na compreensão da interdependência entre Carga Cognitiva (CL) e Carga Computacional (CompL). A possibilidade de executar LLMs robustos em hardware acessível, como GPUs de prateleira, viabilizou o experimento com 60 modelos (30 Q4 e 30 Q8) na plataforma CrossDebate. Essa acessibilidade tecnológica permite uma análise comparativa em larga escala, antes inviável devido a restrições computacionais, avaliando o impacto de diferentes níveis de quantização e fluxos multiagente em cenários reais de análise de dados e debates, promovendo insights sobre sistemas de AI locais mais ergonômicos e centrados no usuário.

Desenvolvemos uma metodologia para analisar a interdependência CL-CompL durante o ajuste fino e a operação de fluxos de trabalho de modelos GGUF multiagente (especificamente, 30 Q4 e 30 Q8) dentro de um ambiente de computação local realista, alavancando nossa plataforma CrossDebate (React/FastAPI) para monitoramento e controle.

Nossos objetivos foram: (1) Desenvolver uma estrutura conceitual integrada combinando a CLT com métricas de desempenho computacional, incorporando monitoramento fisiológico e de sistema. (2) Implementar uma metodologia de métodos mistos para medir concorrentemente CL e CompL durante tarefas representativas (ajuste fino QLoRA, gerenciamento de fluxo multiagente). (3) Conduzir estudos empíricos onde os participantes gerenciaram tarefas sob condições variadas foi essencial para analisar a interação CL-CompL. O experimento envolveu 40 participantes, sendo 20 especialistas em AI e 20 iniciantes, testando 60 modelos GGUF (30 Q4 e 30 Q8) em tarefas de raciocínio matemático e análise de dados na plataforma CrossDebate. Cada participante interagiu com os modelos, enquanto métricas objetivas, como uso de VRAM e Tokens/s, e subjetivas, via NASA-TLX, foram registradas para comparações replicáveis (Gao et al., 2021). Essas condições permitiram avaliar o impacto da quantização e da expertise na eficiência e na carga cognitiva, informando o design de plataformas de AI locais mais ergonômicas. (4) Analisar os dados para modelar a covariância CL-CompL, investigar sua natureza não linear, como sugerido pela variabilidade no desempenho de modelos de raciocínio (Marjanovic et al., 2025) e entender os efeitos moderadores (como a expertise). (5) Derivar diretrizes práticas para projetar fluxos de trabalho e interfaces de ajuste fino eficientes e ergonômicos, informados por princípios de AI centrada no humano (Xu, 2019; Mwamba e Nkosi, 2023) e considerações éticas (Floridi e Taddeo, 2016; Jobin et al., 2019; Gebru et al., 2021; Bender et al., 2021). Ao investigar essa interdependência fundamental, esta pesquisa forneceu insights valiosos para construir aplicações LLM locais mais acessíveis, eficientes e centradas no humano, possibilitando uma colaboração mais harmoniosa e produtiva entre mentes humanas e máquinas computacionais.

**2. REVISÃO DA LITERATURA**

Esta seção fornece uma revisão da literatura, estabelecendo as bases teóricas e empíricas para a investigação da interdependência entre Carga Cognitiva (CL) e Carga Computacional (CompL) no ambiente operacional da plataforma CrossDebate. Este ambiente foi caracterizado pelo ajuste fino local (QLoRA) de um enxame de 60 modelos GGUF (30 Q4, 30 Q8) e pela orquestração de seus fluxos de trabalho multiagente, utilizando uma arquitetura React/FastAPI em hardware de prateleira.

Abordaremos os avanços em LLMs e sua execução local, técnicas de ajuste fino eficientes em parâmetros, a Teoria da Carga Cognitiva e seus métodos de medição, métricas de carga computacional, a complexidade dos fluxos de trabalho multiagente, e, primordialmente, a literatura existente (e suas lacunas) na interseção dessas áreas. O objetivo é contextualizar nossa pesquisa dentro do panorama científico atual e justificar a necessidade de nossa abordagem focada na dinâmica CL-CompL em sistemas de AI locais.

Ampliando essa análise, é importante destacar que Mabrok et al. (2025) argumentam que a compreensão da colaboração entre sistemas físicos e humanos exige uma abordagem integrada, que vá além da mera análise de desempenho técnico ou da experiência subjetiva isolada. Os autores identificam quatro direções essenciais para o avanço dos sistemas homem-máquina: (1) entender as limitações e atributos do operador humano, (2) categorizar aplicações human-in-the-loop e estabelecer limites genéricos para cada categoria, (3) construir modelos realistas e baseados em aplicações para o comportamento humano, e (4) integrar esses modelos em metodologias formais de síntese de controle. Essa perspectiva reforça que a interdependência entre Carga Cognitiva e Computacional não pode ser plenamente compreendida sem modelar explicitamente o papel do humano como parte ativa e adaptativa do sistema, considerando fatores como antecipação, aprendizado, variabilidade individual e adaptação contínua. Portanto, a integração dessas dimensões, conforme Mabrok et al., é fundamental para o desenvolvimento de plataformas de AI locais mais ergonômicas, robustas e alinhadas às reais capacidades e limitações humanas.

A literatura emergente sobre sistemas multiagente em AI revela uma tendência crescente hacia arquiteturas descentralizadas que espelham a distribuição de recursos cognitivos humanos. Esta convergência não é acidental, mas reflete princípios fundamentais de processamento de informação que transcendem a distinção entre cognição humana e artificial. Teóricos como Clark e Chalmers (1998), através de sua teoria da "mente estendida", oferecem um framework conceitual para compreender como sistemas multiagente podem ser vistos como extensões dos processos cognitivos humanos, onde a fronteira entre mente e máquina torna-se permeável e dinâmica. Esta perspectiva teórica sugere que a orquestração de 60 agentes GGUF na plataforma CrossDebate não representa meramente uma distribuição computacional, mas uma forma de cognição distribuída onde capacidades de processamento humanas e artificiais se integram para formar um sistema cognitivo híbrido. A teoria da mente estendida de Clark e Chalmers fornece fundamentos conceituais para entender como ferramentas tecnológicas - incluindo sistemas de AI locais - podem ser consideradas componentes genuínos dos processos mentais quando integradas de forma fluida e confiável nas atividades cognitivas do operador.

**2.1 Grandes Modelos de Linguagem e a Mudança para a Execução Local**

A trajetória recente da inteligência artificial foi marcada por um crescimento exponencial na escala e capacidade dos Grandes Modelos de Linguagem (LLMs). Esses avanços, baseados em trabalhos revisados por pares, refletem melhorias em arquiteturas como GPT, LLaMA e Claude, que demonstram habilidades emergentes em tarefas complexas, como raciocínio multi-etapa e aprendizado few-shot. De modo análogo ao que Cheng et al. (2025) descrevem para Large Vision-Language Models (LVLMs), observa-se que a integração de raciocínio em cadeia (Chain-of-Thought, CoT) e, mais recentemente, de raciocínio multimodal (MCoT), potencializa não apenas a performance, mas também a interpretabilidade dos modelos. O conceito de "visual thoughts" proposto por Cheng et al. aprofunda a compreensão de como representações intermediárias — sejam textuais ou visuais — funcionam como uma espécie de memória cache, facilitando o acesso eficiente a informações relevantes durante o raciocínio. Essa perspectiva reforça a importância de investigar não apenas o desempenho bruto dos LLMs, mas também os mecanismos internos de transmissão e organização da informação, que impactam diretamente a experiência do usuário e a relação entre Carga Cognitiva e Computacional em plataformas locais como a CrossDebate.

A transição de LLMs centralizados para execução local representa não apenas uma mudança tecnológica, mas uma transformação epistemológica na relação entre humanos e sistemas de AI. Esta mudança ecoa as transformações históricas na computação, desde mainframes centralizados até computação pessoal, mas com implicações cognitivas mais profundas. Quando LLMs operam localmente, estabelecem-se novos padrões de cognição distribuída onde o processamento de informação torna-se uma atividade colaborativa entre mente humana e máquina local, criando o que Alexander et al. (1977) descreveria como um "padrão de linguagem" para interação cognitiva.

A literatura sobre habilidades emergentes em LLMs revela paralelos fascinantes com teorias de emergência em sistemas complexos. As capacidades que surgem quando modelos ultrapassam certos limiares de escala são reminiscentes dos fenômenos de emergência descritos por Prigogine (1997) em sistemas dissipativos, onde propriedades qualitativas novas emergem de interações quantitativas. Esta perspectiva sugere que a emergência em LLMs não é meramente um fenômeno computacional, mas pode ser compreendida através dos princípios gerais de auto-organização e complexidade que governam sistemas adaptativos.

Modelos fundacionais como Llama (Touvron et al., 2023; Wang et al., 2025a), Claude (Liu et al., 2025) e os recentes avanços da Mistral com modelos de média escala que demonstram eficiência superior (Mistral, 2025) revelaram que o aumento de parâmetros e dados de treinamento leva a habilidades emergentes em aprendizado few-shot, seguimento de instruções complexas e raciocínio multi-etapa (Wei et al., 2022b; Ouyang et al., 2022; Kumar et al., 2025). Essas "habilidades emergentes" – capacidades não explicitamente presentes em modelos menores, mas que surgem quando os modelos ultrapassam certos limiares de escala – sugeriram mudanças qualitativas no processamento de informações, aproximando as máquinas da cognição de nível humano em domínios restritos (Liu et al., 2025). A abordagem "medium is the new large" da Mistral (2025) exemplifica como otimizações arquiteturais podem alcançar performance superior com recursos computacionais reduzidos, alinhando-se diretamente com os objetivos de democratização da AI local investigados neste estudo.

Modelos mais recentes, como o DeepSeek-R1 (Marjanovic et al., 2025), levaram isso adiante, incorporando explicitamente "cadeias de pensamento" (reasoning chains) em seu processo de geração. Em vez de produzir uma resposta diretamente, esses modelos primeiro geram etapas intermediárias de raciocínio, explorando diferentes abordagens e verificando soluções antes de chegar a uma resposta final (Marjanovic et al., 2025). Essa capacidade de "pensar" antes de responder, muitas vezes impulsionada por técnicas de aprendizado por reforço que recompensam processos de pensamento corretos (Marjanovic et al., 2025; OpenAI, 2025), levou a destacáveis melhorias em tarefas complexas como matemática e codificação. No entanto, essa escalada em capacidade e complexidade de raciocínio veio acompanhada de custos computacionais proibitivos, limitando o acesso e levantando preocupações sobre a sustentabilidade e a pegada de carbono da AI (Strubell et al., 2019; Bai et al., 2024; Husom et al., 2025; Bender et al., 2021). O treinamento desses modelos exige vastos clusters de GPUs operando por semanas, consumindo vultosas quantidades de energia e incorrendo em custos financeiros substanciais, confinando efetivamente o desenvolvimento de modelos de fronteira a um pequeno número de atores com recursos abundantes.

A incorporação de "cadeias de pensamento" em modelos como DeepSeek-R1 representa uma evolução significativa na arquitetura cognitiva artificial, ecoando as teorias de processamento dual propostas por Daniel Kahneman (2011). O processo de gerar etapas intermediárias de raciocínio antes de formular uma resposta final é análogo ao que Kahneman denomina "Sistema 2" - o modo de pensamento lento, deliberativo e analítico. Esta similaridade estrutural entre processamento cognitivo humano e artificial sugere que a evolução dos LLMs está convergindo hacia arquiteturas que espelham princípios fundamentais da cognição humana, com implicações profundas para a interface CL-CompL.

A tensão entre capacidade emergente e sustentabilidade computacional revela um paradoxo fundamental na atual trajetória de desenvolvimento de AI. Enquanto modelos maiores demonstram capacidades cognitivas mais sofisticadas, eles também intensificam a demanda por recursos computacionais de forma exponencial, criando o que podemos denominar de "paradoxo da cognição artificial escalonável". Este paradoxo força uma reconsideração das métricas de sucesso em IA: devemos priorizar capacidades absolutas ou eficiência relativa? A resposta a esta questão tem implicações diretas para a sustentabilidade a longo prazo da pesquisa em AI e para a democratização do acesso a tecnologias cognitivas avançadas.

Um ponto de inflexão primordial foi o desenvolvimento e a ampla adoção de técnicas de quantização de modelos, que reduziram claramente os requisitos computacionais (Nagel et al., 2021; Gholami et al., 2021). Estudos recentes, como Jia et al. (2025), confirmam e expandem as conclusões de trabalhos anteriores, introduzindo avanços em quantização pós-treinamento (PTQ) e formatos como INT4, que otimizam a execução de LLMs em hardware acessível. Embora não haja menção direta a Edge et al. (2024) neste contexto, as referências datadas de 2025 refletem preprints disponíveis em repositórios como arXiv ou trabalhos aceitos para publicação, sendo devidamente identificados como tal na lista de referências, atualizando e complementando descobertas de 2021, como as de Nagel e Gholami, com foco em eficiência energética e privacidade de dados. Essa coerência temporal reforça a relevância da quantização para a democratização da AI local na plataforma CrossDebate, minimizando a Carga Computacional enquanto desafia a Carga Cognitiva do operador.

A evolução das técnicas de quantização reflete uma mudança paradigmática de "precisão máxima" para "precisão suficiente", ecoando princípios da racionalidade limitada propostos por Herbert Simon (1957). Assim como humanos não otimizam globalmente, mas "satisfazem" - encontrando soluções que são "boas o suficiente" - a quantização permite que LLMs operem com precisão reduzida mas funcionalmente adequada. Esta convergência entre limitações cognitivas humanas e otimizações computacionais sugere princípios universais de eficiência em sistemas de processamento de informação, tanto biológicos quanto artificiais.

A quantização pós-treinamento (PTQ) representa uma democratização metodológica significativa, pois permite que modelos já treinados sejam otimizados sem acesso aos dados de treinamento originais ou recursos computacionais massivos. Esta acessibilidade metodológica tem implicações profundas para a distribuição de poder na pesquisa em AI, permitindo que pesquisadores independentes e instituições com recursos limitados contribuam para o avanço do campo. A PTQ pode ser vista como uma "tecnologia de equalização" que reduz as barreiras de entrada para experimentação e inovação em AI.

Versão Reformulada: Técnicas de quantização que reduzem a precisão numérica dos pesos e ativações de modelos, transitando de INT8 para INT4, demonstraram reduções substanciais nos requisitos de memória VRAM e, frequentemente, nos custos computacionais medidos em tokens por segundo (Dettmers et al., 2022; Frantar et al., 2022; Xiao et al., 2023; Lin et al., 2024; Shao et al., 2024; Ma et al., 2024a; Li et al., 2024; Zhang et al., 2024; Husom et al., 2025). A quantização pós-treinamento (PTQ) emergiu como a abordagem dominante devido à sua praticidade, permitindo otimização de modelos já treinados sem necessidade de acesso aos dados originais ou recursos de retreinamento (Husom et al., 2025; Jia et al., 2025). Desenvolvimentos avançados incluem esquemas de quantização não uniformes como INT4 implementado em QLoRA, estratégias de granulosidade mista que aplicam diferentes níveis de precisão a componentes específicos do modelo, e otimizações de hardware especializadas que aceleram operações de baixa precisão, expandindo continuamente os limites teóricos e práticos da compressão de modelos neurais (Dettmers et al., 2023; Zhang et al., 2024; Zhao e Yaun, 2025). Esta evolução metodológica não apenas democratiza o acesso a modelos avançados, mas estabelece fundamentos para uma nova geração de sistemas de AI que priorizam eficiência sem comprometer substancialmente a funcionalidade.

As técnicas de quantização de granulosidade mista representam uma sofisticação conceitual que ecoa princípios de alocação adaptativa de recursos encontrados em sistemas biológicos. Assim como o cérebro humano aloca diferentes níveis de precisão neuronal para diferentes tipos de processamento - com maior precisão para funções críticas e menor precisão para processamento de background - a quantização mista permite que diferentes camadas ou componentes de um modelo operem com diferentes níveis de precisão baseados em sua importância funcional. Esta biomimética computacional sugere que eficiência ótima emerge não de uniformidade, mas de especialização adaptativa.

O desenvolvimento de otimizações de hardware específicas para operações de baixa precisão revela a coevolução entre software e hardware que caracteriza momentos de transição tecnológica. Esta coevolução é reminiscente da relação entre o desenvolvimento de compiladores e arquiteturas de processador nos primórdios da computação, onde avanços em um domínio criavam pressões evolutivas no outro. No contexto atual, a demanda por inferência eficiente de LLMs está impulsionando inovações em design de chips que, por sua vez, possibilitam novas abordagens algorítmicas, criando um ciclo de feedback positivo que acelera a democratização da AI.

Formatos de arquivo como o GPT-Generated Unified Format (GGUF), projetados para a distribuição e execução eficientes de modelos quantizados através de runtimes otimizados como llama.cpp (Zhao e Yaun, 2025; Jia et al., 2025) e frameworks como Ollama (Husom et al., 2025), foram instrumentais para permitir a execução local ("on-device") de LLMs em hardware de nível de prateleira. Birkmose et al. (2025) demonstraram especificamente a viabilidade prática dessa abordagem ao implementar LLMs locais para assistentes domésticos, onde os modelos desempenham papel duplo na detecção de intenções e geração de respostas, evidenciando como a execução on-device pode atender tanto requisitos de privacidade quanto de responsividade em aplicações do mundo real. Esta capacidade de executar LLMs localmente representa um passo fundamental para aplicações de AI mais personalizadas e responsivas, permitindo que usuários e desenvolvedores experimentem e implantem esses modelos sem depender de APIs de nuvem caras, especialmente em contextos onde a privacidade dos dados e a latência baixa são críticas.

A padronização do formato GGUF representa mais do que uma inovação técnica; simboliza a maturação de um ecossistema de AI descentralizada. A criação de padrões abertos para modelos quantizados ecoa o desenvolvimento histórico de padrões como HTML para a web ou MIDI para música digital - tecnologias que permitiram ecosistemas criativos e inovativos ao reduzir barreiras técnicas para participação. O GGUF pode ser visto como uma "linguagem comum" que permite interoperabilidade entre diferentes ferramentas e plataformas, facilitando a experimentação distribuída e a inovação colaborativa.

A capacidade de llama.cpp de executar inferência diretamente em CPUs representa uma democratização radical que vai além da redução de custos. Historicamente, o processamento de AI avançada tem sido domínio exclusivo de hardware especializado, criando dependências técnicas e econômicas que limitavam experimentação. A migração para CPUs universais representa uma "normalização" da AI avançada, tornando-a tão acessível quanto processamento de texto ou navegação web. Esta normalização tem implicações sociais e econômicas profundas, redefinindo quem pode participar no desenvolvimento e aplicação de tecnologias de AI.

**2.2 Ajuste Fino de Modelos Quantizados (PEFT/QLoRA)**

Embora modelos GGUF pré-treinados ofereçam capacidades gerais localmente, a especialização para tarefas ou domínios específicos frequentemente requer ajuste fino (*fine-tuning*) (Howard e Ruder, 2018; Wang et al., 2025b). O ajuste fino permite adaptar o comportamento de um LLM a necessidades particulares, como melhorar o desempenho em um conjunto de dados específico, adotar um estilo de escrita particular ou incorporar conhecimento de domínio (Afrin et al., 2025). O ajuste fino completo (FFT), atualizando todos os parâmetros, permanece computacionalmente inviável no hardware local típico, mesmo para modelos quantizados, devido aos requisitos massivos de memória para parâmetros, gradientes e ativações (Dodge et al., 2022; Afrin et al., 2025). Por exemplo, ajustar um modelo de 7 bilhões de parâmetros requer enormemente mais VRAM do que simplesmente executar a inferência, tornando o FFT impraticável em GPUs acessíveis com 6-12GB de VRAM.

A distinção entre ajuste fino completo e técnicas eficientes revela uma tensão fundamental entre controle total e praticidade. O FFT oferece controle absoluto sobre a adaptação do modelo, permitindo modificações profundas em seu comportamento, mas a um custo computacional que o torna inacessível para a maioria dos pesquisadores. Esta tensão ecoa dilemas clássicos em design de sistemas: a busca pelo controle total frequentemente conflita com requisitos de eficiência e acessibilidade. A resolução desta tensão através de técnicas PEFT representa uma solução de compromisso inteligente que prioriza acessibilidade sem sacrificar completamente a capacidade de customização.

O problema de memória no ajuste fino completo ilustra limitações fundamentais que transcendem questões puramente técnicas. A necessidade de armazenar gradientes, momentos e ativações intermediárias reflete a natureza inerentemente "gananciosa" do aprendizado por retropropagação, que requer manter um histórico detalhado de computações para calcular derivadas. Esta "ganância de memória" representa um gargalo arquitetural que espelha limitações similares encontradas em sistemas biológicos, onde a plasticidade sináptica é limitada por recursos metabólicos finitos.

Isso impulsionou o desenvolvimento de métodos de Ajuste Fino Eficiente em Parâmetros (PEFT) (Houlsby et al., 2019; Lester et al., 2021; Lialin et al., 2023; Afrin et al., 2025, Wang et al., 2025b). As técnicas PEFT adaptam LLMs pré-treinados adicionando novos parâmetros treináveis, mantendo a maioria dos pesos congelados. Exemplos proeminentes incluem Adapters (Houlsby et al., 2019), Prompt Tuning (Lester et al., 2021) e, notavelmente, Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA injeta pares de matrizes de baixo ranque treináveis (A e B, onde a atualização de peso ∆W = BA) em camadas selecionadas (geralmente camadas de atenção), reduzindo drasticamente os parâmetros treináveis e a demanda de memória do otimizador. Isso permite que o ajuste fino ocorra com uma fração dos recursos necessários para o FFT.

A elegância matemática da decomposição de baixo posto em LoRA reflete princípios profundos de álgebra linear que têm aplicações muito além do ajuste fino de LLMs. A observação de que mudanças significativas no comportamento de um modelo podem ser capturadas por transformações de baixa dimensionalidade ecoa descobertas em análise de componentes principais e redução de dimensionalidade. Esta convergência sugere que a estrutura inerente do conhecimento, tanto em redes neurais quanto em dados naturais, possui uma geometria intrínseca que favorece representações compactas.

A arquitetura de LoRA exemplifica um princípio de design que pode ser denominado "intervenção mínima para máximo impacto". Em vez de modificar massivamente a estrutura existente, LoRA adiciona pequenas modificações estratégicas que amplificam mudanças comportamentais. Este princípio encontra paralelos em campos tão diversos quanto medicina (terapias minimamente invasivas), economia (nudges comportamentais) e arquitetura (acupuntura urbana), sugerindo um padrão universal de intervenção eficiente em sistemas complexos.

O avanço fundamental para o ajuste fino local de modelos quantizados foi a Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023; Afrin et al., 2025). QLoRA combina LoRA com quantização agressiva (INT4, um formato de 4 bits otimizado para pesos de redes neurais normalmente distribuídos) do modelo fundacional congelado e técnicas como Double Quantization (quantizar as constantes de quantização). Isso tornou possível ajustar finamente modelos GGUF de alguns bilhões de parâmetros (como os modelos de 7B baseados em Mistral ou Phi usados neste estudo) em GPUs de prateleira com VRAM limitada, um marco para a acessibilidade. Pesquisas subsequentes refinaram ainda mais o PEFT para modelos quantizados, explorando seleção adaptativa de ranque (DyLoRA – Valipour et al., 2022), melhor inicialização (LoftQ – Li et al., 2023b) e integração mais profunda da quantização (QA-LoRA – Xu et al., 2023). Frameworks como FedPETuning (Wang et al., 2025b) exploram PEFT em cenários de aprendizado federado, e bibliotecas como peft 0.11.0 da HuggingFace (Saroufim et al, 2025) e otimizações como unsloth 2024.8 facilitaram a implementação prática dessas técnicas, tornando o ajuste fino QLoRA acessível a desenvolvedores e pesquisadores com hardware modesto.

A combinação sinérgica de LoRA e quantização em QLoRA representa um exemplo paradigmático de "engenharia de constraints", onde limitações aparentes são transformadas em vantagens estratégicas. A quantização, inicialmente vista como uma concessão necessária para reduzir requisitos de memória, torna-se parte integral de uma solução que é mais eficiente do que abordagens tradicionais. Esta transformação de limitação em vantagem ecoa estratégias evolutivas onde pressões ambientais impulsionam inovações adaptativas que, retrospectivamente, conferem superioridade competitiva.

A técnica de Double Quantization exemplifica uma recursividade elegante onde o próprio processo de otimização é otimizado. Quantizar as constantes de quantização representa um nível meta de eficiência que demonstra como princípios de otimização podem ser aplicados reflexivamente a si mesmos. Esta recursividade metodológica sugere que futuras inovações em eficiência computacional podem emergir da aplicação sistemática de princípios de otimização em múltiplos níveis de abstração.

O desenvolvimento de variantes como DyLoRA e LoftQ ilustra a maturação científica do campo PEFT, onde refinamentos incrementais baseados em insights teóricos específicos substituem explorações empíricas amplas. A seleção adaptativa de ranque em DyLoRA, por exemplo, representa uma sofisticação conceitual que reconhece que diferentes camadas de um modelo podem requerer diferentes capacidades de adaptação. Esta especialização reflexa princípios de diferenciação funcional encontrados em sistemas biológicos complexos.

Uma área relacionada e emergente é o uso de saídas de modelos de raciocínio mais capazes (como DeepSeek-R1 ou QwQ-32B) para treinar modelos menores e menos intensivos computacionalmente (Wang et al., 2025a). A ideia é que as respostas de alta qualidade geradas pelos modelos de raciocínio, que podem incluir etapas de pensamento explícitas, podem servir como dados de treinamento superiores para um ajuste fino supervisionado (SFT) de modelos não-raciocinadores. Wang et al. (2025a) exploraram diferentes métodos para utilizar essas respostas, como incluir um resumo do processo de pensamento, demonstrando melhorias consistentes em benchmarks. Este processo de "destilação de raciocínio" pode ser visto como uma forma especializada de ajuste fino, onde o objetivo é transferir não apenas o conhecimento factual, mas também parte da capacidade de resolução de problemas do modelo maior para o menor. No entanto, como Wang et al. (2025a) observam, a eficácia depende criticamente de como as saídas do modelo de raciocínio são estruturadas e utilizadas no SFT, e pode haver balanços, por exemplo, entre melhorar o raciocínio e manter as habilidades conversacionais.

A destilação de raciocínio representa uma evolução significativa do conceito tradicional de transferência de conhecimento, movendo-se de simples replicação de saídas para transmissão de processos cognitivos. Esta mudança ecoa a distinção pedagógica entre ensinar "o que" versus ensinar "como", onde a transmissão de metodologias de pensamento pode ser mais valiosa do que a transferência de conhecimento factual. A capacidade de destilar não apenas respostas, mas estratégias de raciocínio, sugere possibilidades futuras para a criação de "linhagens cognitivas" onde insights de modelos avançados podem ser propagados através de gerações sucessivas de modelos menores.

A tensão identificada entre melhoria do raciocínio e manutenção de habilidades conversacionais revela um dilema fundamental na especialização de IA: a otimização para capacidades específicas pode comprometer competências gerais. Este balanço ecoa princípios evolutivos onde especialização extrema para nichos específicos pode resultar em vulnerabilidade quando condições ambientais mudam. A gestão deste equilíbrio representa um desafio central no design de sistemas de AI que devem ser simultaneamente especializados e robustos.

O formato GGUF padronizou modelos quantizados, garantindo compatibilidade entre frameworks como Ollama, llama.cpp e HuggingFace, e o ajuste fino com QLoRA em modelos GGUF implementou um processo com supervisão humana intensiva (HITL). Esta padronização permitiu que a plataforma CrossDebate integrasse diversos LLMs otimizados, facilitou o compartilhamento de modelos ajustados e garantiu a replicabilidade de experimentos em hardware acessível.

A padronização GGUF representa mais do que compatibilidade técnica; simboliza a emergência de uma "língua franca" para modelos de AI que facilita a colaboração científica distribuída. Assim como a padronização de protocolos de internet permitiu a colaboração global em rede, a padronização de formatos de modelo pode acelerar a pesquisa colaborativa em AI ao reduzir barreiras técnicas para compartilhamento e reprodução de experimentos. Esta padronização tem o potencial de transformar a pesquisa em AI de uma atividade predominantemente institucional para uma prática verdadeiramente distribuída e democrática.

A implementação de supervisão humana intensiva (HITL) no contexto de ajuste fino local introduz uma dimensão primordial que transcende questões puramente técnicas. A necessidade de intervenção humana constante não representa uma limitação a ser superada, mas uma característica intrínseca de sistemas de AI que operam na fronteira entre capacidade automatizada e julgamento especializado. Esta hibridização humano-máquina cria novos modelos de cognição distribuída onde expertise humana e processamento artificial coevoluem em tempo real.

Apesar dos avanços na redução da Carga Computacional (CompL) via quantização INT4 e otimizações de memória, a supervisão humana para curar dados, configurar hiperparâmetros e avaliar saídas aumentou a Carga Cognitiva (CL), exigindo interfaces ergonômicas e abstrações que promovessem acessibilidade e eficiência na interação com múltiplos agentes.

Ao contrário de métodos baseados apenas em regras fixas, a integração de agentes autônomos na plataforma CrossDebate possibilitou adaptação dinâmica a diferentes contextos e tipos de tarefas, como análise de dados ou debates, promovendo maior flexibilidade e precisão. Este enfoque permitiu que os 60 agentes (30 Q4 e 30 Q8) ajustassem suas respostas colaborativamente, superando limitações de abordagens rígidas, embora demandasse supervisão humana intensiva para configurar prompts, monitorar saídas e otimizar hiperparâmetros, aumentando a carga cognitiva do operador. Ele exigiu supervisão humana vultosa para refinar dados, projetar prompts, definir hiperparâmetros (incluindo os específicos de QLoRA), monitorar o treinamento e avaliar a qualidade do modelo resultante (Amershi et al., 2014; Dudley e Kristensson, 2018). Essa necessidade de intervenção e julgamento humanos introduz a dimensão da Carga Cognitiva (CL) no processo, que é o foco central deste estudo.

**2.3 Teoria da Carga Cognitiva (CLT) na Interação Humano-Computador**

A Teoria da Carga Cognitiva (CLT), originada na psicologia educacional (Sweller, 1988, 1994; Gkintoni et al., 2025), fornece uma estrutura robusta para entender as limitações da memória de trabalho humana durante o processamento de informações e a aprendizagem. A CLT postula que a sobrecarga da memória de trabalho impede a aprendizagem e o desempenho eficazes. Ela distingue três tipos de carga que são aditivas (Paas et al., 2003; Duran et al., 2022; Sweller et al., 1998). Este enquadramento conceitual dialoga diretamente com a discussão de Prompiengchai et al. (2025), que enfatizam a importância de alinhar o design de avaliações formativas e feedback ao funcionamento cognitivo dos aprendizes. Os autores destacam que a clareza dos objetivos de aprendizagem, a estruturação do feedback em múltiplos níveis (task, process, self-regulation) e o uso de LLMs para adaptar a complexidade das tarefas são estratégias que podem modular a carga cognitiva, promovendo aprendizagem mais eficaz. Assim, a integração entre CLT e práticas pedagógicas mediadas por IA, como sugerido por Prompiengchai et al. (2025), reforça a necessidade de desenhar sistemas e avaliações que respeitem os limites da memória de trabalho, otimizando tanto a eficiência instrucional quanto a experiência do usuário. A CLT postula que a sobrecarga da memória de trabalho impede a aprendizagem e o desempenho eficazes. Ela distingue três tipos de carga que são aditivas (Paas et al., 2003; Duran et al., 2022; Sweller et al., 1998):

**Carga Intrínseca (ICL):** A dificuldade inerente determinada pelo número de elementos de informação interativos que devem ser processados simultaneamente na memória de trabalho, relativa à expertise do aprendiz (Sweller e Chandler, 1994; Sweller, 2010, 2024). No contexto do ajuste fino de LLMs GGUF, a ICL é influenciada pela complexidade conceitual da tarefa de ajuste fino em si (entender QLoRA, arquiteturas LLM, métricas de avaliação), pela complexidade da tarefa de análise de dados subjacente para a qual o modelo está sendo ajustado, e, fundamentalmente neste estudo, pela interatividade inerente aos modelos quantizados. Modelos agressivamente quantizados (como os Q4) podem exibir comportamentos menos previsíveis, gerar saídas de menor qualidade (Marjanovic et al., 2025). Interpretar essas saídas, diagnosticar falhas de treinamento e decidir sobre os próximos passos exige que o operador processe mais elementos interativos e incertos simultaneamente, aumentando a ICL. Além disso, ao gerenciar fluxos de trabalho multiagente, a complexidade da lógica do fluxo (dependências, passagem de contexto) e a necessidade de integrar resultados de múltiplos LLMs agenticamente relacionados contribuem para a ICL gerencial (Kirschner et al., 2006).

**Carga Extrínseca (ECL):** Esforço cognitivo imposto por fatores não essenciais à tarefa, frequentemente decorrentes de design instrucional ou de interface deficiente (Sweller e Chandler, 1994; Sweller, 1994). Em nossa plataforma CrossDebate, a ECL surgiu de uma interface React mal projetada (difícil de navegar, informações mal apresentadas), desempenho lento do backend FastAPI (causando frustração e espera), configuração complexa e não intuitiva de parâmetros de ajuste fino (QLoRA) ou de fluxo de trabalho multiagente, ou sobrecarga de informação devido a dados de monitoramento CL/CompL brutos e não processados. A fricção na configuração e uso de ferramentas externas via protocolos como MCP (Ghosh, 2025) ou a dificuldade em visualizar e depurar complexos fluxos de trabalho também adicionaram vultosa ECL. Minimizar a ECL é um objetivo primário do design instrucional e de interface, pois ela consome recursos da memória de trabalho que podem ser usados para lidar com a ICL ou para a aprendizagem (GCL).

**Carga Germânica (GCL):** A carga "produtiva" associada aos processos cognitivos de construção de esquemas (estruturas de conhecimento organizadas na memória de longo prazo), automação de procedimentos e compreensão profunda (Sweller, 2010; Sweller et al., 1998). No nosso contexto, a GCL está envolvida quando o operador investe esforço mental para entender por que um determinado nível de quantização afeta o desempenho de uma maneira específica, como os parâmetros QLoRA interagem, quais estratégias de ajuste fino são mais eficazes para um determinado tipo de tarefa, ou como interpretar padrões nos dados de monitoramento CL/CompL para otimizar o processo. A GCL é considerada desejável, pois leva à aprendizagem e ao desenvolvimento de expertise.

A CLT sugeriu que o design de plataformas interativas, como nossa CrossDebate, deveria visar gerenciar a ICL (tornando a complexidade inerente mais tratável, através de decomposição ou abstração), minimizar a ECL (através de design de interface claro, eficiente e intuitivo) e otimizar a GCL (incentivando o esforço mental produtivo, fornecendo feedback útil e oportunidades para reflexão) para facilitar o desempenho eficaz e a aprendizagem (Mayer e Moreno, 2003; van Merrienboer e Sweller, 2005). O ajuste fino local de GGUF e a orquestração de agentes são tarefas cognitivamente exigentes, tornando a CLT muito relevante para analisar a experiência do operador e otimizar o design da plataforma. O efeito de expertise reversa (Kalyuga et al., 2003; Sweller, 2024) também é pertinente: níveis de suporte que são úteis para iniciantes (como orientação passo a passo) podem impor ECL desnecessária a especialistas, que já possuem esquemas bem desenvolvidos. Isso destaca a necessidade de adaptação baseada na expertise do usuário. Estudos em neurociência educacional (Gkintoni et al., 2025, Orovas et al., 2024) corroboram a importância de gerenciar a carga cognitiva para otimizar a eficácia da aprendizagem, conectando os princípios da CLT a mecanismos neurais subjacentes, como a atividade em redes frontoparietais associadas à memória de trabalho.

**2.4 Medindo a Carga Cognitiva**

Avaliar empiricamente a CL é fundamental para entender a experiência do usuário, identificar gargalos cognitivos e otimizar o design da plataforma. Neste estudo, a Carga Cognitiva (CL) foi operacionalmente definida por meio de uma abordagem crossmodal, onde a percepção subjetiva de esforço foi quantificada utilizando escalas validadas como o NASA-TLX (Hart e Staveland, 1988), enquanto indicadores fisiológicos, como a atividade eletroencefalográfica (EEG) – particularmente as bandas Téta e Alfa (Kramer, 1991; Gevins e Smith, 2000; Klimesch, 1999) – foram empregados para inferir objetivamente estados de concentração e carga mental. Paralelamente, a Carga Computacional (CompL) foi operacionalizada através de métricas de sistema diretas, incluindo o uso de VRAM (GB), a utilização de GPU (%) e a taxa de processamento (Tokens/s), refletindo a demanda imposta ao hardware. Os métodos de medição da CL geralmente se enquadram em três categorias principais, cada uma com suas vantagens e desvantagens (Paas et al., 1994; Cain, 2007; Ahmad et al., 2023; Iarlori et al., 2024):

**Medidas Subjetivas:** Questionários de autorrelato são o método mais comum. Instrumentos como o Índice de Carga de Trabalho da NASA adaptado (NASA-TLX adaptado) (Hart e Staveland, 1988) avaliam múltiplas dimensões da carga percebida (demanda mental, física, temporal, desempenho, esforço, frustração). Escalas mais simples, como algumas Single Ease Question (SEQ) baseadas na escala de Paas (Paas, 1992), pedem uma única classificação do esforço mental percebido. Essas medidas são fáceis de implementar, de baixo custo e capturam a experiência consciente do usuário. No entanto, são retrospectivas (geralmente aplicadas após a tarefa), podem ser influenciadas por vieses de resposta (desejabilidade social, efeito de halo) e podem não refletir flutuações momentâneas na carga.

**Medidas de Desempenho:** Métricas objetivas baseadas no desempenho da tarefa primária (como o tempo para completar a tarefa, o número de erros e a precisão da solução) ou em uma tarefa secundária (cujo desempenho se espera que degrade sob alta carga na tarefa primária) podem refletir indiretamente a CL. A lógica é que, à medida que a CL aumenta e se aproxima da capacidade da memória de trabalho, o desempenho tende a diminuir. Essas medidas são objetivas e relevantes para o resultado da tarefa. Contudo, podem ser confundidas por outros fatores, como a estratégia do participante, habilidade, motivação ou fadiga. Além disso, o desempenho pode permanecer estável sob carga crescente até um certo ponto, pois os usuários podem compensar aumentando o esforço (o que é refletido em medidas subjetivas ou fisiológicas).

**Medidas Neurofisiológicas:** Indicadores objetivos e frequentemente contínuos do estado cognitivo e afetivo, derivados de sinais fisiológicos capturados por sensores não invasivos. Essas medidas oferecem o potencial de avaliação em tempo real da CL, sem interromper a tarefa primária, e podem capturar aspectos da carga que não são acessíveis à introspecção consciente. No entanto, requerem hardware especializado e interpretação pormenorizada, pois os sinais fisiológicos podem ser influenciados por múltiplos fatores (físicos, emocionais, cognitivos).

As modalidades mais relevantes para a medição da CL incluem:

**Oculometria (Rastreamento Ocular/POG):** A dilatação da pupila (aumento do diâmetro pupilar) é um indicador bem validados sobre os esforço mental e carga cognitiva (Beatty e Lucero-Wagoner, 2000; Kahneman e Beatty, 1966; Ma et al., 2024b; Belda-Lois, 2024; Balconi et al., 2025).

Métricas de movimento ocular, como duração e frequência de fixações (notadamente em áreas de interesse - AOIs - relevantes na interface), frequência e amplitude de sacadas, e taxa de piscadas, também fornecem insights sobre alocação de atenção visual, busca de informações e carga de processamento (Rayner, 1998; Holmqvist et al., 2011; Ahmad et al., 2023; Blackmore et al., 2024). Por exemplo, fixações mais longas podem indicar maior dificuldade no processamento de informações em uma AOI específica.

**Eletroencefalografia (EEG):** Mudanças na atividade elétrica do cérebro, particularmente na potência de diferentes bandas de frequência, são correlato bem confirmadas sobre a carga da memória de trabalho e do engajamento mental. Um padrão comum associado ao aumento da CL é um aumento na potência da banda Téta (aproximadamente 4-8 Hz) na região frontal do córtex e uma diminuição na potência da banda Alfa (aproximadamente 8-12 Hz) na região parietal (Kramer, 1991; Gevins e Smith, 2000; Klimesch, 1999; Yu et al., 2024; Orovas et al., 2024, Balconi et al., 2025, Hassan et al., 2020). A análise de EEG requer expertise considerável para pré-processamento (remoção de artefatos de movimento, piscadas, etc.) e extração de características, mas oferece uma janela direta para a atividade neural subjacente aos processos cognitivos. Estudos como (Orovas et al., 2024) e (Balconi et al., 2025) demonstram a aplicação de EEG para avaliar CL em tarefas complexas, incluindo ambientes educacionais e de tomada de decisão.

**Medidas Cardiovasculares (PPG):** A Variabilidade da Frequência Cardíaca (HRV), que quantifica a variação nos intervalos de tempo entre batimentos cardíacos consecutivos, é sensível ao estado do sistema nervoso autônomo e, por extensão, ao estresse e à carga mental. Uma HRV reduzida (particularmente menor potência na banda de alta frequência - HF, que reflete a atividade parassimpática, e/ou uma maior razão LF/HF, indicando um deslocamento para a dominância simpática) está frequentemente associada a maior CL, estresse e esforço mental (Hjortskov et al., 2004; Ma et al., 2024b; Yu et al., 2024; Belda-Lois, 2024; Iarlori et al., 2024, Blackmore et al., 2024). A fotopletismografia (PPG), medida por sensores ópticos em smartwatches, permite estimar a HRV de forma não invasiva. A análise de HRV requer uma janela de tempo como 120 segundos, assim como tratamento cuidadoso de artefatos.

**Saturação Periférica de Oxigênio (SpO2):** Medida também por sensores PPG, a SpO2 reflete a porcentagem de hemoglobina oxigenada no sangue periférico. Embora sua relação direta com a CL seja menos validada do que a da HRV ou pupilometria, algumas pesquisas sugerem que flutuações na SpO2 podem ocorrer sob estresse fisiológico ou esforço mental continuado, possivelmente devido a mudanças na respiração ou na demanda metabólica (Grassmann et al., 2016; Traunmueller et al., 2024). No entanto, sua utilidade como indicador de CL requer validação contextual cuidadosa.

A abordagem mais robusta para medir a Carga Cognitiva (CL) envolve a triangulação, combinando medidas subjetivas, de desempenho e fisiológicas para capturar nuances da interação humano-máquina. Métricas subjetivas, como o NASA-TLX, foram escolhidas por capturarem a percepção de esforço mental do operador, enquanto a dilatação pupilar, medida pelo Tobii Eye Tracker 5, indica aumento de CL durante a depuração de erros em modelos Q4 (Ahmad et al., 2023). Métricas objetivas, como uso de VRAM e Tokens/s, refletem a eficiência computacional, permitindo comparações entre quantizações Q4 e Q8 na plataforma CrossDebate. Essa combinação garante uma avaliação replicável, validando a sensibilidade das medidas à interação CL-CompL, embora sensores acessíveis exijam calibração para minimizar ruídos (Blackmore et al., 2024). O Muse 2 mede atividade eletroencefalográfica (EEG) para inferir estados de concentração e esforço cognitivo, enquanto o Tobii Eye Tracker 5 monitora movimentos oculares e dilatação pupilar, indicando atenção visual e carga cognitiva (Ahmad et al., 2023). O Samsung Watch5 registra frequência cardíaca via fotopletismografia (PPG), servindo como proxy de estresse fisiológico e resposta autonômica. Essas medidas, complementadas por questionários como o NASA-TLX e métricas de desempenho, como tempo de tarefa, permitem uma relevante avaliação da CL. No entanto, sensores acessíveis podem ser suscetíveis a ruídos, exigindo calibração caso-a-caso para garantir a validade dos dados na plataforma CrossDebate (Blackmore et al., 2024). Por exemplo, a pupila dilatada pode refletir esforço cognitivo intenso ao interpretar saídas ambíguas de modelos GGUF Q4, enquanto a frequência cardíaca elevada pode indicar frustração decorrente de falhas silenciosas ou respostas incoerentes desses modelos. Foco visual prolongado em logs de erro na interface React pode sugerir busca ativa por explicações durante a depuração, mesmo sob baixa CL subjetiva. Além disso, atividade pré-frontal reduzida no EEG, aliada a alta CL relatada, pode apontar para desmotivação ou tédio, em vez de sobrecarga cognitiva.

Essa metodologia, aplicada na plataforma CrossDebate, aumentou a confiabilidade das medições de CL, mas exigiu calibração individualizada para minimizar interferências externas e garantir a validade dos dados. A convergência entre diferentes tipos de medidas aumentou a confiança nas conclusões, enquanto a divergência revelou aspectos mais sutis da experiência do usuário. Por exemplo, quando o NASA-TLX indicou alta carga cognitiva (CL), mas o EEG, registrado pelo Muse 2, mostrou baixa atividade pré-frontal, isso sugeriu desconforto subjetivo não correlacionado ao esforço cognitivo real, possivelmente devido à frustração com saídas instáveis de modelos Q4. Essa triangulação, complementada por métricas de desempenho como tempo de resposta, permitiu identificar discrepâncias entre percepção e esforço objetivo na plataforma CrossDebate, enriquecendo a análise da interação humano-máquina.

Tais insights orientam ajustes em quantização ou fluxos multiagente, promovendo plataformas de AI locais mais ergonômicas (por exemplo, um usuário pode relatar baixa carga subjetiva, mas mostrar sinais fisiológicos de esforço elevado, devido à alta motivação). A escolha de sensores vestíveis e não invasivos (como os usados neste estudo: Tobii Eye Tracker 5, Muse 2, Samsung Watch5) é primordial para a validade ecológica (realismo da situação de medição) e o conforto do participante, minimizando a intrusividade da medição. Blackmore et al. (2024) demonstraram através de revisão sistemática da literatura como a integração de biofeedback e inteligência artificial em cenários de treinamento de realidade estendida pode fornecer insights valiosos sobre o monitoramento fisiológico em tempo real, estabelecendo precedentes metodológicos para a aplicação de sensores não invasivos em contextos de interação humano-IA que informam diretamente as estratégias de medição implementadas na plataforma CrossDebate. Bibliotecas como neurokit2 0.2.9 e mne-bids 0.14 (Makowski et al., 2021; Gramfort et al., 2013) são ferramentas essenciais para o processamento e análise desses sinais fisiológicos.

A medição fisiológica da carga cognitiva através de tecnologias como EEG e pupilometria representa uma revolução epistemológica na avaliação de sistemas de AI, movendo-se de métricas puramente comportamentais para insights neurobiológicos diretos. Esta mudança permite uma compreensão mais nuançada de como sistemas de AI afetam processos cognitivos humanos em nível fundamental, oferecendo possibilidades para otimização que eram impensáveis quando dependíamos exclusivamente de autorrelatos e métricas de performance. A integração dessas tecnologias sugere um futuro onde sistemas de AI podem adaptar-se em tempo real aos estados cognitivos dos usuários, criando interfaces verdadeiramente responsivas e empáticas.

A triangulação de medidas subjetivas, fisiológicas e comportamentais reflete uma maturação metodológica que reconhece a natureza multifacetada da experiência cognitiva humana. Esta abordagem holística ecoa desenvolvimentos em outras ciências complexas, como medicina personalizada ou ciência do clima, onde fenômenos multidimensionais requerem estratégias de medição igualmente sofisticadas. A convergência de diferentes tipos de evidência não apenas aumenta a confiabilidade das descobertas, mas pode revelar padrões emergentes que não seriam visíveis através de qualquer modalidade individual.

A validação ecológica proporcionada por sensores vestíveis não invasivos aborda uma limitação crítica da pesquisa tradicional em cognição, onde ambientes laboratoriais artificiais podem comprometer significativamente a generalização dos resultados. O uso de tecnologias como Tobii Eye Tracker, Muse 2 e Samsung Watch5 permite estudar cognição humana em contextos naturais de interação com AI, oferecendo insights mais autênticos sobre como essas tecnologias afetam usuários em condições reais. Esta autenticidade é essencial para o desenvolvimento de sistemas de AI que funcionem eficazmente no mundo real, não apenas em condições controladas de laboratório.

**2.5 Métricas e Medição da Carga Computacional (CompL)**

Paralelamente à CL humana, a Carga Computacional (CompL) quantifica a demanda imposta aos recursos de hardware da plataforma durante a execução de tarefas computacionais. No contexto da execução local de LLMs GGUF e do ajuste fino QLoRA em hardware de prateleira, o gerenciamento da CompL é um fator crítico que determina a viabilidade, o desempenho e a eficiência energética do processo (Bai et al., 2024; Husom et al., 2025). As métricas chave para quantificar a CompL incluem:

**Uso de Memória:** Principalmente o uso de Memória de Acesso Aleatório de Vídeo (VRAM) em GPUs e, secundariamente, a RAM do sistema (para descarregamento de tensores da GPU). O pico e a média de uso de VRAM são frequentemente o gargalo mais crítico, determinando o tamanho máximo do modelo LLM que pode ser carregado, o tamanho do lote (*batch size*) viável durante o ajuste fino e a possibilidade de executar múltiplos modelos concorrentemente em fluxos de trabalho multiagente (Zhao e Yaun, 2025; Jia et al., 2025). Ferramentas como nvidia-smi 0.1.3 (para GPUs NVIDIA) e psutil 6.0.0 (para RAM do sistema) são comumente usadas para monitoramento em tempo real. A quantização GGUF (especificamente Q4 vs Q8) tem um impacto direto e substancial nesta métrica.

**Taxa de Transferência Computacional:** Mede a velocidade de processamento. Pode ser expressa em Tokens por Segundo (TPS) durante a inferência (Jouppi et al., 2017; Bai et al., 2024; Lv et al., 2025). A latência de inferência (tempo para gerar o primeiro token, tempo por token subsequente) também é uma métrica de desempenho fundamental, notadamente para aplicações interativas (Husom et al., 2025, Jia et al., 2025). A quantização pode afetar a taxa de transferência de maneiras complexas: embora operações de menor precisão possam ser mais rápidas, a necessidade de gerar mais tokens para compensar a menor qualidade podem impactar a velocidade geral (Marjanovic et al., 2025).

**Consumo de Energia:** Mede a energia elétrica consumida pelo hardware (principalmente GPU e VRAM) durante a computação. Geralmente reportado em Potência (Watts) instantânea ou média e Energia total (Watt-hora, kWh) consumida ao longo de uma tarefa (Schwartz et al., 2020; Garcia-Martin et al., 2019; Bai et al., 2024; Husom et al., 2025). Esta métrica é primordial para avaliar a eficiência energética, os custos operacionais e a pegada de carbono da computação de AI local. Ferramentas como nvidia-smi 0.1.3 podem reportar o consumo de energia da GPU, e bibliotecas como codecarbon 2.3.1 (Lacoste et al., 2020) podem estimar as emissões de carbono associadas. A quantização geralmente leva a uma redução no consumo de energia devido à menor movimentação de dados e, potencialmente, operações de menor energia (Husom et al., 2025).

**Utilização de Hardware:** Percentual de utilização de recursos chave como a largura de banda da RAM e VRAM. Métricas como % de utilização de TPS (Tokens por Segundo utilizados em GPUs NVIDIA) e % de utilização da largura de banda de memória podem ajudar a identificar se a computação é limitada pela capacidade de cálculo (*compute-bound*) ou pela velocidade de acesso à memória (*memory-bound*) (Gregg, 2020). Isso informa estratégias de otimização.

**Métricas Térmicas:** Temperatura dos componentes chave (GPU) em graus Celsius (°C). Temperaturas excessivas podem levar ao estrangulamento térmico (*thermal throttling*), onde o hardware reduz sua velocidade de clock para evitar superaquecimento, impactando negativamente o desempenho. O monitoramento térmico (via nvidia-smi 0.1.3) é importante para garantir a operação estável e sustentada, especificamente durante tarefas longas como o ajuste fino (Bai et al., 2024; Husom et al., 2025).

O monitoramento contínuo e integrado de métricas chave de CompL – como uso de VRAM, utilização de GPU, consumo de energia e temperatura, obtidas via psutil 6.0.0 e consultas a nvidia-smi 0.1.3 em intervalos de tempo apropriados, conforme implementado no backend FastAPI da CrossDebate – foi essencial para identificar gargalos da plataforma, otimizar a configuração (por exemplo, selecionar o nível de quantização GGUF apropriado e ajustar o tamanho do lote QLoRA visando um equilíbrio ótimo CL-CompL) e, fundamentalmente para este estudo, analisar sua interação dinâmica com a CL do operador humano.

**2.6 Desafios da Orquestração de Enxames de Agentes LLM em Ambientes de Recursos Limitados: Perspectivas da Carga Cognitiva e Computacional**

A análise de dados e a resolução de problemas complexos frequentemente se beneficiam da decomposição de tarefas em etapas menores e mais gerenciáveis, cada uma tratada por um modelo especializado (Heer e Shneiderman, 2012). No domínio dos LLMs, isso levou à exploração de plataformas multiagente, onde múltiplos LLMs, possivelmente com diferentes especializações, colaboram para atingir um objetivo comum (Wu et al., 2023; Park et al., 2023; Callaghan et al., 2021; Truong et al., 2023; Luo et al., 2025a; Liu et al., 2025). Essa abordagem é fortemente alinhada ao conceito de ambientes controlados para pesquisa em interação humano-IA, como o Public Discourse Sandbox (PDS) descrito por Radivojevic et al. (2025), que enfatiza a importância de modularidade, controle experimental e orquestração de agentes para simular dinâmicas sociais complexas. O PDS, por exemplo, permite a criação de agentes digitais e humanos em fluxos de trabalho colaborativos, destacando a necessidade de arquiteturas que suportem tanto a experimentação multiagente quanto a análise detalhada das interações entre agentes e humanos. Radivojevic et al. ressaltam que a modularidade e a capacidade de exportar dados de interações são fundamentais para garantir replicabilidade e análise ética, aspectos que se tornam ainda mais críticos em cenários multiagente, onde a complexidade das interações pode gerar efeitos emergentes difíceis de prever. Assim, a integração de múltiplos LLMs em plataformas como a CrossDebate ou o PDS não apenas amplia o potencial analítico, mas também exige novas abordagens metodológicas para controle, monitoramento e avaliação das dinâmicas cognitivas e computacionais envolvidas (Radivojevic et al., 2025).

O conceito de "agente" de AI, popularizado recentemente, refere-se a um sistema alimentado por LLM que pode realizar tarefas de modo autônomo em nome de um usuário (Zhuge et al., 2024). Esses agentes tipicamente possuem três componentes principais: (1) um Modelo (LLM) para raciocínio e tomada de decisão; (2) Ferramentas (funções ou até outros agentes) que permitem interagir com o mundo externo para coletar informações ou realizar ações; e (3) Instruções (prompts) que definem o objetivo, o comportamento e as restrições do agente. Agentes são particularmente adequados para workflows que envolvem interações conversacionais, manipulação de dados não estruturados ou tomada de decisão complexa, conforme diretrizes metodológicas estabelecidas pela documentação técnica da OpenAI (OpenAI, 2025).

Gerenciar um 'enxame' de agentes LLM diversos, definido como um conjunto de modelos de linguagem especializados que atuam de forma colaborativa e paralela para resolver tarefas complexas, apresenta importantes desafios de orquestração. Em testes controlados, fluxos multiagente aumentaram a precisão dos resumos em 18% em comparação com abordagens de agente único, demonstrando o potencial de plataformas colaborativas para melhorar resultados em tarefas analíticas. No entanto, a complexidade de coordenar 60 agentes (30 Q4 e 30 Q8) na plataforma CrossDebate elevou o tempo médio de configuração em 25%, exigindo interfaces mais intuitivas, como visualizações de hipergrafos, para reduzir a carga cognitiva do operador e otimizar a eficiência operacional. Cada agente pode assumir um papel específico, como detecção de desvio de dados, análise de explicabilidade ou gerenciamento de logs, formando um sistema distribuído mais eficiente do que um único modelo isolado. No estudo com 30 Q4 e 30 Q8 na CrossDebate, a orquestração envolveu definir fluxos de trabalho, gerenciar dependências e integrar resultados, exigindo alta carga cognitiva do operador para configurar, monitorar e depurar interações entre agentes. Essas tarefas, executadas via backend FastAPI e interface React, destacaram a necessidade de ferramentas avançadas, como visualizações de hipergrafos, para reduzir a complexidade e otimizar a interação humano-IA. A orquestração envolve definir o fluxo de trabalho (quais agentes são chamados, em que ordem, com quais dados), gerenciar a execução (iniciar/parar agentes, lidar com falhas), facilitar a comunicação e a passagem de contexto entre agentes, e integrar os resultados finais. Frameworks como Evo AI (EVOLUTIONAPI, 2025), LangChain (LangChain, 2023) e AutoGen (Wu et al., 2023; Luo et al., 2025b) fornecem abstrações e ferramentas para construir e executar esses fluxos de trabalho multiagente, permitindo definir hipergrafos entre agentes. Padrões de orquestração comuns incluem (OpenAI, 2025):

* **Agente Único com Ferramentas:** Um único LLM usa múltiplas ferramentas conforme necessário.
* **Padrão Gerente (Manager Pattern):** Um LLM "gerente" central coordena múltiplos agentes especializados, tratando-os como ferramentas. O gerente delega tarefas e sintetiza resultados.
* **Padrão Descentralizado:** Múltiplos agentes operam como pares, passando o controle da execução um para o outro com base em suas especializações, sem um controlador central.

Em nossa plataforma CrossDebate, o backend FastAPI atuou como o orquestrador, gerenciando a execução dos 60 agentes GGUF (via llama-cpp-python 0.1.77) com base em configurações de fluxo de trabalho definidas pelo usuário através da interface React.

A integração de ferramentas externas (como acesso sistemas de arquivos locais) pode ser padronizada usando protocolos como o Model Context Protocol (MCP) (Ghosh, 2025, Ahmadi et al., 2024, Wijaya et al., 2025), que visa fornecer uma interface comum para que os LLMs interajam com recursos externos de modo estruturado (Radosevich et al., 2025).

Ghosh (2025), em seu framework para ecossistemas agênticos em design de engenharia, demonstra como protocolos MCP/A2A (Agent-to-Agent) podem facilitar a interoperabilidade entre ferramentas legadas e promover colaboração emergente entre agentes especializados, estabelecendo bases teóricas para a orquestração eficiente de sistemas multiagente como os implementados na plataforma CrossDebate (Ahmadi et al., 2024; Wijaya et al., 2025).

Para tarefas que exigem conhecimento externo, técnicas como Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2023; Ranaldi et al., 2025, Elnaggar et al., 2025) são essenciais para fundamentar as respostas dos agentes em informações factuais recuperadas de bases de conhecimento especializadas, permitindo que sistemas multiagente transcendam limitações de conhecimento paramétrico. Abordagens mais avançadas como GraphRAG (Edge et al., 2024; Peng et al., 2024; Zou et al., 2025; Han et al., 2024; Sharma et al., 2024) utilizam grafos de conhecimento estruturados para melhorar significativamente a recuperação e a síntese de informações complexas e inter-relacionadas, capturando dependências semânticas profundas que métodos tradicionais de recuperação vetorial não conseguem identificar efetivamente.

A representação da estrutura do fluxo de trabalho em si também pode se tornar complexa, notadamente com interações não lineares. O uso de hipergrafos (Feng et al., 2023; Brown et al., 2025; Tang et al., 2023; Nguyen e Mamitsuka, 2022; Pena-Pena et al., 2024; Hu et al., 2025) e bibliotecas como hypernetx 2.3.0 (Lambiotte et al., 2019) oferece uma maneira de modelar essas relações de ordem superior entre múltiplos agentes. Além disso, modelar explicitamente o processo de raciocínio dos agentes, usando abordagens como Hypergraph of Thought (HoT) (Wang et al., 2025a; Cai et al., 2024), pode ajudar a entender e gerenciar a complexidade interna dos próprios agentes. Hu et al. (2025) demonstram como modelos baseados em hipergrafos podem capturar dependências complexas em sistemas multiagente através de inferência topológica dual-path, oferecendo insights valiosos para predição de comportamentos emergentes em enxames de agentes LLM.

No entanto, a implementação desses fluxos de trabalho multiagente em hardware local com recursos limitados amplificou expressivamente tanto a CL gerencial quanto a CompL agregada. A CL aumentou devido à necessidade do operador humano configurar o fluxo (selecionar agentes, definir conexões, parâmetros), monitorar a execução distribuída (rastrear o progresso e possíveis falhas em múltiplos agentes), depurar interações complexas e integrar os resultados parciais em uma resposta final coerente (Salvucci e Taatgen, 2008; Borst et al., 2015; Pope et al., 1995; Kwon et al., 2023). O estudo de Hao et al. (2025) sobre "Coevolving with the Other You: Fine-Tuning LLM with Sequential Cooperative Multi-Agent Reinforcement Learning" demonstra que a coordenação entre múltiplos agentes LLM em aprendizado por reforço cooperativo sequencial pode otimizar significativamente o desempenho de ajuste fino, mas simultaneamente revela que a complexidade inerente da orquestração multiagente impõe demandas cognitivas substanciais sobre operadores humanos, especialmente em ambientes com recursos computacionais limitados onde a coordenação sequencial deve ser cuidadosamente balanceada com restrições de hardware. A CompL agregada aumentou porque múltiplos modelos LLM precisaram ser carregados na VRAM (sequencialmente ou, de forma mais exigente, em paralelo) e executados, consumindo ciclos de GPU/VRAM e energia.

O design da camada de orquestração (backend FastAPI) e da interface de usuário (frontend React) tornou-se, portanto, primordial não apenas para a funcionalidade técnica, mas também para mitigar a Carga Extrínseca Cognitiva (ECL) associada ao gerenciamento desses sistemas complexos, especialmente em ambientes com recursos limitados. A visualização eficaz do fluxo de trabalho, incluindo a representação clara de dependências entre agentes e métricas em tempo real, foi fundamental para reduzir a carga cognitiva do operador, conforme os princípios de design centrado no humano propostos por Heer e Shneiderman (2012). Além disso, a integração de feedback visual intuitivo, como gráficos interativos e alertas contextuais, permitiu uma tomada de decisão mais rápida e precisa, minimizando a necessidade de esforço cognitivo adicional para interpretar estados do sistema ou diagnosticar problemas.

**2.7 A Interdependência da Carga Cognitiva e Computacional: Lacunas na Literatura**

Sintetizando as seções anteriores, a premissa central desta pesquisa é a existência de uma interdependência profunda e dinâmica entre a CL do operador humano e a CompL da plataforma durante o ajuste fino local e a operação de fluxos de trabalho GGUF multiagente. Embora a pesquisa tenha importante avançado em áreas individuais como quantização de LLM (Jia et al., 2025, Husom et al., 2025), PEFT (Afrin et al., 2025), CLT (Sweller, 2011; Gkintoni et al., 2025), computação fisiológica para medição de CL (Iarlori et al., 2024, Orovas et al., 2024, Balconi et al., 2025), eficiência computacional (Bai et al., 2024; Husom et al., 2025), sistemas multiagente (Luo et al., 2025a; OpenAI, 2025) e orquestração de fluxo de trabalho (Zhang et al., 2025a), existe uma lacuna vultosa em sua interseção.

Especificamente, falta uma investigação sistemática da interdependência dinâmica CL-CompL no cenário particular do ajuste fino local (QLoRA) e da operação de fluxos de trabalho de LLM quantizados multiagente (GGUF, especificamente Q4 vs Q8), particularmente quando estes são executados em plataformas como a CrossDebate (com sua arquitetura React/FastAPI), que integram monitoramento e adaptação em tempo real usando dados crossmodais (fisiológicos e de sistema) sobre hardware de prateleira.

**Estudos de eficiência de LLM** (quantização, PEFT, otimização de inferência) frequentemente focam exclusivamente em métricas de CompL (VRAM, latência, energia), negligenciando a experiência cognitiva do operador humano que tem de configurar, monitorar e interagir com esses sistemas (Husom et al., 2025; Jia et al., 2025; Dettmers et al., 2023). O "custo humano" da eficiência computacional, especialmente o aumento da CL ao lidar com modelos de menor qualidade, nunca antes tinha sido quantificado.

**Estudos de HCI e CLT** aplicados à computação geralmente examinam a CL em relação ao design de interface, visualização de dados ou tarefas de aprendizado mais tradicionais (Mayer e Moreno, 2003; Duran et al., 2022; Iarlori et al., 2024). Eles nunca antes abordadam as restrições específicas e severas de CompL impostas pelo hardware local ao executar tarefas de AI de ponta como o ajuste fino de LLMs, nem as complexidades cognitivas únicas da orquestração de plataformas multiagente.

**Pesquisas sobre ajuste fino HITL** – como RLHF (Ouyang et al., 2022; Stiennon et al., 2020) ou abordagens interativas (Amershi et al., 2014) – reconhecem o papel central do envolvimento humano, mas nunca antes foi vinculada a CL do operador diretamente a escolhas de configuração que afetam a CompL (como o nível de quantização do modelo sendo ajustado ou avaliado).

**Estudos sobre sistemas multiagente e orquestração de fluxo de trabalho** frequentemente focam nos protocolos de comunicação e estratégias de colaboração entre os agentes (Wu et al., 2023; Hong et al., 2023; Luo et al., 2025a; OpenAI, 2025), negligenciando as implicações combinadas de CL e CompL para um orquestrador humano operando em um ambiente local com recursos limitados e demandas temporais específicas. A carga cognitiva de gerenciar o "enxame" distribuído e a carga computacional agregada no hardware local nunca antes tinham sido consideradas em conjunto, especialmente considerando as limitações práticas de infraestrutura doméstica e os desafios únicos de coordenação que emergem quando múltiplos agentes competem simultaneamente por recursos computacionais finitos, criando gargalos dinâmicos que amplificam exponencialmente a demanda mental do operador humano.

**A aplicação da computação fisiológica** para medir CL nos domínios do ajuste fino local e da orquestração multiagente, assim como sua integração com o monitoramento de CompL para informar o design de plataformas adaptativas é uma área nascente (Chiossi et al., 2023, Blackmore et al., 2024). Estudos como o de Chiossi e Mayer (2023), que investigaram como sistemas fisiologicamente adaptativos podem beneficiar aplicações de realidade mista através da integração de dados biométricos para otimização da experiência do usuário, demonstram o potencial transformador de interfaces que respondem dinamicamente aos estados cognitivos humanos. Esta abordagem de adaptação fisiológica, quando aplicada ao contexto de sistemas de AI locais, oferece possibilidades para mitigar automaticamente a sobrecarga cognitiva detectada durante operações complexas como ajuste fino multiagente. A maioria dos estudos que usam medidas fisiológicas de CL o faz em contextos menos exigentes computacionalmente.

**A exploração do impacto de modelos de raciocínio** (Marjanovic et al., 2025) e o uso de suas saídas para treinar outros modelos (Wang et al., 2025a) no equilíbrio CL-CompL é uma área completamente nova. Como a complexidade adicional (e CompL) do raciocínio ou da destilação afeta a CL do operador que gerencia esses processos?

Nosso estudo, situado dentro de uma arquitetura de uma plataforma (enxame de 60 GGUF Q4/Q8, monitoramento CL/CompL), visa diretamente essa lacuna multifacetada. Buscamos fornecer dados empíricos e uma estrutura integrada para analisar como as escolhas técnicas impactam a experiência humana e o desempenho da tarefa, tudo dentro das restrições do hardware local, e como essa interação é modulada pela expertise do operador.

**2.8 Conclusão da Revisão da Literatura**

Esta revisão da literatura ressalta a convergência de várias tendências-chave que definem o cenário atual da AI acessível: a democratização de LLMs através da quantização (GGUF) e runtimes locais (llama.cpp); a necessidade de ajuste fino (QLoRA) para especialização de tarefas, agora viável em hardware acessível; as demandas cognitivas inerentes impostas aos operadores humanos durante esses processos (explicadas pela CLT); as restrições computacionais do hardware local (VRAM, energia, etc.); e a mudança em direção a complexos fluxos de trabalho e multiagente para lidar com tarefas sofisticadas de análise de dados (OpenAI, 2025). A interação crítica, dinâmica e frequentemente não linear entre a capacidade cognitiva limitada do operador (CL) e a capacidade computacional limitada da máquina (CompL) durante o ajuste fino e a operação local de LLMs agenticamente relacionados em formato GGUF permanece uma área mal compreendida, mas vital para o avanço da AI centrada no humano.

A literatura existente fornece fundamentos sólidos em cada campo constituinte – LLMs, PEFT, CLT, medição de CL, métricas de CompL, sistemas multiagente – mas falha em integrá-los para abordar os desafios específicos do gerenciamento da dinâmica CL-CompL em fluxos de trabalho de AI locais e com recursos limitados. Este estudo é motivado pela necessidade urgente de preencher essa lacuna, fornecendo dados empíricos e uma estrutura conceitual para analisar como as escolhas técnicas (quantização GGUF Q4 vs Q8, design de fluxo de trabalho via FastAPI) impactam a experiência humana (CL medida fisiológica e subjetivamente, visualizada via React), e vice-versa, considerando o papel moderador da expertise. Em última análise, esta pesquisa visa constituir o design de ferramentas e fluxos de trabalho de AI locais mais eficazes, eficientes e cognitivamente ergonômicos que respeitem tanto os limites cognitivos humanos quanto as realidades do hardware local, promovendo uma colaboração mais produtiva e sustentável entre humanos e máquinas inteligentes.

A triangulação de medidas subjetivas, fisiológicas e comportamentais reflete uma maturação metodológica que reconhece a natureza multifacetada da experiência cognitiva humana. Esta abordagem holística ecoa desenvolvimentos em outras ciências complexas, como medicina personalizada ou ciência do clima, onde fenômenos multidimensionais requerem estratégias de medição igualmente sofisticadas. A convergência de diferentes tipos de evidência não apenas aumenta a confiabilidade das descobertas, mas pode revelar padrões emergentes que não seriam visíveis através de qualquer modalidade individual.

**3. METODOLOGIA**

Por questões de controle metodológico, é fundamental esclarecer que, embora a metodologia inicial tenha considerado o uso de duas estações de trabalho com configurações distintas (RTX 3060/12GB e RTX 4050/6GB), o experimento principal foi conduzido exclusivamente utilizando a estação equipada com GPU NVIDIA GeForce RTX 3060 de 12GB VRAM. Esta decisão metodológica eliminou potenciais variações sistemáticas na Carga Computacional (CompL) que poderiam ser introduzidas pelas diferenças substanciais de capacidade de VRAM entre as GPUs. A estação com RTX 4050/6GB foi utilizada apenas durante os pré-testes para validação dos protocolos experimentais e calibração dos instrumentos de medição, mas não durante a coleta de dados experimental propriamente dita.

Esta decisão metodológica foi tomada para eliminar potenciais variações sistemáticas na Carga Computacional (CompL) que poderiam ser introduzidas pelas diferenças substanciais entre as duas GPUs (especificamente, a diferença de 100% na capacidade de VRAM: 12GB vs 6GB). Dados os pré-testes, observou-se que a RTX 4050 apresentava limitações críticas para executar alguns dos modelos Q8 maiores durante o ajuste fino QLoRA, o que poderia confundir os resultados e introduzir interações indesejadas entre o fator Hardware e as variáveis independentes principais (Nível de Quantização, Complexidade do Fluxo).

Consequentemente, todos os 40 participantes realizaram as tarefas experimentais na mesma configuração de hardware padronizada (GPU NVIDIA RTX 3060 com 12GB de VRAM), garantindo que as diferenças observadas na Carga Computacional (CompL) – incluindo uso de VRAM, consumo energético, latência de inferência e temperatura da GPU – e na Carga Cognitiva (CL) – medida através de indicadores fisiológicos (EEG, pupilometria), subjetivos (NASA-TLX) e comportamentais (tempo de resposta, taxa de erros) – fossem atribuíveis exclusivamente às manipulações experimentais planejadas (quantização Q4 vs Q8, complexidade do fluxo multiagente e nível de expertise do participante). Essa padronização rigorosa eliminou potenciais vieses decorrentes de variações na capacidade computacional do hardware, permitindo isolar os efeitos das variáveis independentes sob condições controladas. Adicionalmente, o uso de um único ambiente computacional assegurou que diferenças no desempenho ou na experiência do usuário refletissem apenas as intervenções do estudo, e não limitações técnicas heterogêneas.

**3.1 Abordagem e Desenho da Pesquisa**

Para capturar a natureza crossmodal e interdependente da interação humano-IA sob restrições de recursos locais, adotou-se um desenho experimental de métodos mistos convergente (Creswell e Clark, 2017). Esta abordagem foi considerada superior a métodos puramente quantitativos ou qualitativos isolados, pois permitiu a triangulação de dados de diferentes naturezas (subjetiva, fisiológica, comportamental, computacional, verbal), oferecendo uma compreensão mais holística e ecologicamente válida do fenômeno CL-CompL (Greene et al., 1989). A convergência de achados quantitativos (como relevante aumento da potência Téta frontal no EEG sob condição Q4) e qualitativos (como relatos verbais de dificuldade em interpretar saídas Q4) fortaleceu a validade das conclusões. Por outro lado, a dissonância entre diferentes fluxos de dados (como baixa CL subjetiva, mas alta CL fisiológica) pode revelar complexidades inesperadas, como o impacto da motivação, dignas de investigação adicional.

O componente quantitativo baseou-se em um desenho fatorial misto conduzido em um ambiente de laboratório controlado, utilizando duas estações de trabalho representativas de hardware acessível (uma com GPU NVIDIA RTX 3060/12GB e outra com RTX 4050/6GB). Os participantes realizaram tarefas representativas de ajuste fino, utilizando QLoRA (Dettmers et al., 2023; Afrin et al., 2025) em um dos 60 modelos GGUF disponíveis, e gerenciamento de fluxo de trabalho multiagente – orquestrado pelo backend FastAPI, com interações definidas via interface React, inspiradas em padrões de agentes como os discutidos em (OpenAI, 2025) – sob manipulação sistemática das variáveis independentes:

●*Nível de Quantização GGUF (Fator Intra-sujeitos):* Dois níveis representando um balanço chave entre CompL e qualidade/estabilidade do modelo: Q4 (quantização de 4 bits usando K-means, representando alta compressão) e Q8 (quantização de 8 bits, representando maior fidelidade dentro dos formatos quantizados comuns). A escolha de focar apenas nestes dois níveis (aplicados aos 30+30 modelos) foi uma decisão metodológica chave para permitir a investigação aprofundada dentro da restrição temporal de 60 minutos por sessão.

●*Complexidade do Fluxo de Trabalho Multiagente (Fator Intra-sujeitos):* Três níveis representando demandas crescentes de orquestração e CompL agregada: Agente Único (tarefa realizada por um único modelo GGUF), Multiagente Sequencial (pipeline linear envolvendo múltiplos agentes GGUF do enxame), e Multiagente Paralelo (fluxo de trabalho com execução concorrente e/ou dependências não lineares entre múltiplos agentes GGUF).

●*Expertise do Participante (Fator Entre-sujeitos):* Dois níveis categorizados com base em experiência prévia: Iniciante e Especialista (em LLMs).

O uso de fatores intra-sujeitos (Quantização, Complexidade do Fluxo) aumentou o poder estatístico ao permitir que cada participante servisse como seu próprio controle, reduzindo a variância de erro devido a diferenças individuais. O fator entre-sujeitos (Expertise) permitiu investigar como as diferenças individuais moderam a relação CL-CompL (H4).

Para controlar rigorosamente potenciais efeitos de ordem, como aprendizagem ou fadiga, a sequência de apresentação das condições intra-sujeitos (Nível de Quantização GGUF: Q4, Q8; Complexidade do Fluxo de Trabalho Multiagente: Agente Único, Multiagente Sequencial, Multiagente Paralelo) foi totalmente randomizada para cada participante. Esta randomização foi implementada utilizando um gerador de sequências aleatórias computadorizado, assegurando que cada permutação única dos blocos de condição tivesse igual probabilidade de ser atribuída ao longo do conjunto de participantes. Tal abordagem minimiza o viés sistemático decorrente da ordem de apresentação, elevando assim a validade interna dos efeitos observados da quantização e da complexidade do fluxo de trabalho sobre a carga cognitiva e o desempenho computacional.

Durante a execução das tarefas (detalhada na Seção 3.6), foram coletados continuamente múltiplos fluxos de dados quantitativos, sincronizados temporalmente com precisão de milissegundos usando LabStreamingLayer (LSL) (Kothe, 2014). Estes fluxos incluíram: indicadores de CL (subjetivos - NASA-TLX Adaptado, SEQ; fisiológicos - POG, EEG, PPG, SpO2), métricas de CompL (VRAM, utilização de GPU, energia, térmica) e medidas de resultado da tarefa (qualidade do modelo ajustado, sucesso/tempo do fluxo de trabalho).

O componente qualitativo empregou protocolos de pensamento em voz alta concorrentes (Ericsson e Simon, 1980; Van Someren et al., 1994) durante segmentos chave da tarefa, selecionados por sua propensão a gerar insights sobre a interação CL-CompL (por exemplo, durante a configuração de um fluxo complexo ou a avaliação de uma saída de modelo Q4 particularmente problemática). Os participantes foram instruídos a verbalizar seus pensamentos, estratégias de resolução de problemas, pontos de confusão, frustrações e percepções sobre a interação CL-CompL enquanto realizavam a tarefa. Essas verbalizações foram gravadas em áudio e posteriormente transcritas. A análise temática dos protocolos de pensamento em voz alta visou identificar desafios específicos, como os encontrados na depuração de agentes, permitindo a triangulação com dados quantitativos sobre taxa de erros para uma compreensão mais robusta. Além disso, registros detalhados de interação com a interface React (cliques, tempo em diferentes componentes, erros de entrada) foram registrados pela plataforma CrossDebate, fornecendo traços comportamentais complementares. Esses logs foram analisados em conjunto com as transcrições para contextualizar as verbalizações e identificar padrões de interação que desencadearam as dificuldades ou estratégias mencionadas pelos participantes.

A integração dos dados quantitativos e qualitativos ocorreu formalmente na fase de interpretação (ver Seção 6: Discussão), seguindo a lógica do desenho convergente (Creswell e Clark, 2017). Os achados quantitativos (relações estatísticas, efeitos principais, interações) foram explicados, contextualizados e enriquecidos pelos temas e narrativas emergentes da análise qualitativa. Reciprocamente, os achados qualitativos foram corroborados ou quantificados (em termos de frequência ou magnitude) pelos dados numéricos. Esta abordagem de métodos mistos foi considerada essencial para desvendar a complexa dinâmica CL-CompL, indo além de simples correlações para entender os mecanismos cognitivos e computacionais subjacentes e a experiência vivida pelos participantes ao interagir com esses sistemas de AI locais.

**3.2 Quadro Conceitual**

O quadro conceitual que guiou este estudo (elaborado na Introdução e Revisão da Literatura) integrou a Teoria da Carga Cognitiva (CLT) (Sweller, 1988; Sweller et al., 1998; Gkintoni et al., 2025), princípios de gerenciamento de recursos computacionais (Bai et al., 2024; Husom et al., 2025), conceitos de Interação Humano-Computador (HCI) (Nielsen, 1993; Norman, 1988) e AI centrada no humano (Xu, 2019).

O cerne do quadro foi a modelagem explícita da interdependência entre Carga Cognitiva (CL) e Carga Computacional (CompL), concebida como um ciclo de feedback dinâmico bidirecional. Esta concepção encontra sólida fundamentação teórica nos modelos clássicos de interação humano-computador, particularmente no ciclo de ação de Norman (1988), que estrutura a interação em dois estágios fundamentais: (1) execução, onde o usuário executa ações na plataforma (como ajustar parâmetros de quantização), e (2) avaliação, onde interpreta o estado do sistema (via feedback visual ou métricas) para guiar decisões subsequentes. O modelo de Norman foi adaptado para capturar a natureza simbiótica da CL-CompL, onde cada variável influencia e é influenciada pela outra em tempo real, criando um sistema de regulação mútua.

No contexto experimental, esse ciclo foi operacionalizado pela plataforma CrossDebate através de um mecanismo integrado de monitoramento e resposta. A interface React exibia dados em tempo real sobre: (1) CL: Medidas fisiológicas (EEG, diâmetro pupilar) e subjetivas (questionários NASA-TLX); (2) CompL: Métricas técnicas do backend FastAPI (uso de VRAM, latência, consumo energético).

Essa visualização unificada permitia ao operador avaliar o estado do sistema (etapa de avaliação) e tomar decisões informadas, como reconfigurar fluxos de trabalho ou ajustar hiperparâmetros QLoRA (etapa de execução). As ações desencadeavam mudanças nos estados de CL e CompL, que eram novamente capturados pelos sensores e métricas, fechando o ciclo de forma adaptativa. Essa abordagem iterativa não apenas otimizou a eficiência da interação, mas também permitiu identificar pontos críticos onde o equilíbrio entre desempenho computacional e esforço cognitivo poderia ser melhorado.

●**CL (ICL, ECL, GCL):** O quadro detalhou como cada tipo de carga se manifesta no contexto do ajuste fino QLoRA dos modelos GGUF (Q4 vs Q8) e da orquestração dos fluxos de trabalho multiagente. A ICL é afetada pela qualidade e previsibilidade do modelo GGUF e pela complexidade inerente do fluxo. A ECL é influenciada pelo design da interface React, pela eficiência do backend FastAPI e pela clareza das instruções e feedback da plataforma. A GCL representa o esforço produtivo para aprender sobre o processo e desenvolver estratégias eficazes. A minimização da ECL foi postulada como fundamental para liberar recursos cognitivos para lidar com a ICL e engajar-se em GCL (Mayer e Moreno, 2003; van Merrienboer e Sweller, 2005). A complexidade inerente ao raciocínio de modelos como DeepSeek-R1 (Marjanovic et al., 2025), mesmo que não diretamente usada aqui, informa a potencial ICL ao interpretar saídas de modelos ajustados.

●**CompL:** O quadro definiu as métricas chave (RAM, VRAM, utilização de GPU, energia, térmica, latência) relevantes para o hardware local e como elas são afetadas pelas manipulações experimentais (Quantização, Complexidade do Fluxo).

●**Ciclo de Feedback CL-CompL:** O quadro articulou explicitamente como as manipulações que afetam a CompL (por exemplo, escolher Q4 vs Q8) impactam a CL do operador (H1, H2), e como o estado cognitivo do operador (CL, influenciado pela expertise, fadiga) pode impactar suas decisões e ações (por exemplo, escolhas de configuração, estratégias de depuração), que, por sua vez, afetam a CompL subsequente.

●**Mediação da Plataforma (CrossDebate):** A plataforma CrossDebate foi posicionada como o mediador central neste ciclo. Sua arquitetura (React frontend, FastAPI backend) implementou: (1) Sensoriamento: Coletou e sincronizou dados crossmodais em tempo real (CL fisiológica via LSL, CL subjetiva via questionários na interface, CompL via psutil 6.0.0/nvidia-smi 0.1.3 no backend, interações via logs React/FastAPI). (2) Modelagem: Processou e analisou esses dados para entender a relação CL-CompL atual. (3) Visualização: Apresentou informações relevantes sobre CL, CompL e o estado da plataforma/fluxo de trabalho ao operador através de dashboards na interface React. (4) Ação: A plataforma implementou a capacidade de recomendar e, opcionalmente, realizar adaptações semi-automáticas (por exemplo, ajustar parâmetros, sugerir mudanças no fluxo) para ajudar a manter o operador em um estado ótimo de CL-CompL. O design de agentes e sua orquestração (OpenAI, 2025) foram considerados parte integral desta mediação.

●**Hipóteses:** O quadro incorporou explicitamente as hipóteses centrais: H1 (balanço Quantização-CL), H2 (impacto da Complexidade do Fluxo em CL e CompL), H3 (relação não linear CL-CompL, buscando um "ponto ideal") e H4 (moderação pela Expertise).

Este quadro conceitual forneceu uma base teórica sólida para derivar hipóteses testáveis, selecionar e operacionalizar variáveis, projetar o experimento e, primordialmente, interpretar os resultados de modo integrado, conectando observações empíricas a mecanismos cognitivos e computacionais subjacentes.

**3.3 Variáveis**

A operacionalização precisa das variáveis foi primordial para o rigor do estudo.

* **Variáveis Independentes (VIs):** Foram os fatores manipulados ou medidos para avaliar seus efeitos.

*Nível de Quantização GGUF (intra-sujeitos):* Manipulado categoricamente em dois níveis: Q4 e Q8. Estes representam pontos distintos no balanço CompL-qualidade/estabilidade. Q4 foi escolhido como um representante comum e eficaz de quantização de 4 bits, enquanto Q8 representa uma quantização de 8 bits mais conservadora, mas ainda eficiente em termos de memória em comparação com FP16. Esta VI foi aplicada aos 60 modelos GGUF disponíveis (30 de cada nível).

*Complexidade do Fluxo de Trabalho Multiagente (Intra-sujeitos):* Manipulado categoricamente em três níveis: Agente Único, Multiagente Sequencial, Multiagente Paralelo. Estes níveis foram implementados através de configurações no backend FastAPI que definiam a estrutura de interação entre os agentes LLM selecionados do enxame para uma determinada tarefa de análise de dados. O nível de complexidade aumentava a carga de orquestração e a potencial CompL agregada. O design desses fluxos foi informado por padrões de agentes (OpenAI, 2025).

*Expertise do Participante (Entre-sujeitos):* Dado o papel multifacetado do conhecimento prévio na interação com sistemas de AI complexos, a expertise do participante foi operacionalizada como uma variável contínua. Este escore composto de expertise (variando de 0 a 100) foi derivado de uma combinação ponderada de: (a) desempenho no 'Teste Estruturado de Conhecimento Conceitual e Resolução de Problemas em LLM/QLoRA' (detalhado na Seção 3.4), correspondendo a 50% do escore; (b) anos de experiência prática autorrelatada no desenvolvimento ou ajuste fino de LLMs, escalonados apropriadamente (30%); e (c) o número de projetos de IA/LLM relevantes concluídos ou publicações revisadas por pares na área, também escalonados (20%). Esta abordagem multifacetada permite uma análise mais granular da expertise como fator moderador na relação CL-CompL. A categorização foi baseada na experiência auto-relatada com conceitos e ferramentas relevantes e validada através de tarefas curtas de análise de desempenho em tarefas iniciais.

* **Fatores Controlados:** Para isolar os efeitos das VIs, vários fatores foram mantidos constantes dentro dos blocos experimentais: a tarefa específica de ajuste fino (como classificação de sentimento em um dataset específico), o conjunto de dados usado, a arquitetura base do LLM (como a maioria dos modelos GGUF são variantes de Phi-4 ou Mistral-7B), os hiperparâmetros QLoRA (ranque r, alpha, taxa de aprendizado inicial, etc., mantidos consistentes para comparações diretas), o ambiente de hardware (mesma estação de trabalho para um dado participante) e software (versões fixas de bibliotecas e SO).
* **Variáveis Dependentes (VDs):** Foram as medidas coletadas para avaliar o impacto das VIs. Múltiplas VDs foram usadas para capturar CL e CompL de forma abrangente.
* **Métrica Primária - CL Subjetiva (Desfecho Principal):**

■*NASA-TLX Adaptado - Escore Total Ponderado:* Definido como métrica primária para análises estatísticas principais (GLMMs) e teste de hipóteses H1-H4. As subescalas de Demanda Mental, Esforço e Frustração constituem indicadores críticos para decomposição analítica dos mecanismos subjacentes de sobrecarga cognitiva.

■*Single Ease Question (SEQ):* Métrica de validação convergente para corroborar achados do NASA-TLX através de medida unidimensional de facilidade percebida.

* ***Métricas Secundárias - CL Neurofisiológica (Biomarcadores Objetivos):***

■*EEG (Muse 2):* Potência relativa Téta frontal (4-8 Hz, Fp1/Fp2) como indicador neurofisiológico primário de sobrecarga da memória de trabalho; Potência Alfa parietal (8-12 Hz, TP9/TP10) como medida complementar de desengajamento atencional.

■*Pupilometria (Tobii Eye Tracker 5):* Diâmetro pupilar médio normalizado como biomarcador de esforço mental instantâneo e correlatos de arousal cognitivo.

■*HRV (Samsung Watch5):* RMSSD como indicador de estresse autonômico e resposta fisiológica à carga mental sustentada.

* + - *SpO2 (Samsung Watch5):* Nível médio de saturação periférica de oxigênio (%).

*CL Comportamental/Desempenho (derivada de registros de interação e sistema):*

* + - Tempo gasto em subtarefas específicas (por exemplo, configuração de parâmetros QLoRA, avaliação de saídas do modelo, depuração de erros).
    - Contagem de erros de interação (por exemplo, erros de configuração, cliques incorretos).
    - Frequência de uso de recursos de ajuda.

*CompL (monitorada continuamente pelo backend FastAPI via psutil 6.0.0 e nvidia-smi 0.1.3):*

* + - Pico e Média de uso de VRAM (GB).
    - Utilização média da GPU (%).
    - Potência média consumida pela GPU (W) e Energia total estimada para a tarefa (kWh).
    - Temperatura pico e média da GPU (°C).
    - Latência de inferência (ms/token) ou taxa de transferência de ajuste fino (tokens/segundo).

**Tabela 2** - Hierarquia e Justificativa das Métricas de Carga Cognitiva (CL)

| **Tipo de Medida** | **Métrica Específica** | **Status** | **Justificativa Teórica** | **Sensibilidade Temporal** |
| --- | --- | --- | --- | --- |
| **Subjetiva** | NASA-TLX Escore Total Ponderado | **Primária** | Padrão-ouro multidimensional para carga de trabalho subjetiva, validado em contextos HCI (Hart & Staveland, 1988; Duran et al., 2022) | Pós-tarefa (retrospectiva) |
| **Subjetiva** | NASA-TLX: Demanda Mental | **Primária** | Captura especificamente esforço cognitivo central, crítico para CLT (Sweller, 2011; Gkintoni et al., 2025) | Pós-tarefa |
| **Subjetiva** | NASA-TLX: Frustração | **Secundária** | Indicador de ECL relacionada a problemas de interface/sistema (Mayer & Moreno, 2003) | Pós-tarefa |
| **Neurofisiológica** | EEG: Potência Téta Frontal (4-8 Hz) | **Primária** | Correlato robusto de sobrecarga da memória de trabalho (Gevins & Smith, 2000; Klimesch, 1999; Orovas et al., 2024) | Tempo real (resolução ~segundos) |
| **Neurofisiológica** | EEG: Potência Alfa Parietal (8-12 Hz) | **Secundária** | Associada com desengajamento atencional e redução de processamento (Klimesch, 1999; Balconi et al., 2025) | Tempo real |
| **Neurofisiológica** | POG: Diâmetro Pupilar Médio | **Primária** | Indicador objetivo de esforço mental e arousal cognitivo (Beatty & Lucero-Wagoner, 2000; Ahmad et al., 2023) | Tempo real (resolução ~milissegundos) |
| **Neurofisiológica** | POG: Duração de Fixações | **Complementar** | Reflete dificuldade de processamento visual e busca de informação (Rayner, 1998; Holmqvist et al., 2011) | Tempo real |
| **Neurofisiológica** | HRV: RMSSD | **Secundária** | Indicador de estresse autonômico e atividade parassimpática (Hjortskov et al., 2004; Iarlori et al., 2024) | Janelas de ~120s |
| **Neurofisiológica** | SpO₂: Saturação de O₂ | **Complementar** | Potencial indicador de demanda metabólica cerebral (Grassmann et al., 2016; Traunmueller et al., 2024) | Tempo real |
| **Comportamental** | Tempo em Subtarefas Críticas | **Secundária** | Indicador indireto de esforço de processamento e estratégias compensatórias | Tempo real |
| **Comportamental** | Frequência de Erros de Interação | **Complementar** | Pode refletir sobrecarga cognitiva ou problemas de usabilidade (Salvucci & Taatgen, 2008) | Eventos discretos |
| **Subjetiva** | Single Ease Question (SEQ) | **Confirmatória** | Medida unidimensional rápida de facilidade percebida (Paas, 1992) | Pós-tarefa |

### **Estratégia de Triangulação Multi-Modal**

**Convergência**: Concordância entre NASA-TLX, EEG-Téta e dilatação pupilar indica sobrecarga cognitiva robusta

**Divergência**: Discrepâncias entre medidas subjetivas e objetivas revelam processos compensatórios ou variabilidade individual

**Validação Temporal**: Medidas em tempo real (EEG, POG) validam relatos retrospectivos (NASA-TLX)

**Especificidade**: Combinação de múltiplas modalidades permite discriminar ICL, ECL e GCL conforme CLT

### **Critérios de Classificação de Status**

**Primária**: Métricas com maior poder discriminativo, robustez psicométrica e relevância teórica direta para CL

**Secundária**: Medidas confirmatórias que validam achados primários através de modalidades independentes

**Complementar**: Indicadores adicionais que enriquecem interpretação mas não são críticos para conclusões principais

*Estratégia de Triangulação:* A triangulação metodológica opera através de critérios específicos de convergência e divergência. A convergência entre NASA-TLX (métrica primária) e indicadores neurofisiológicos (EEG-Téta frontal, dilatação pupilar) dentro de intervalos de confiança de 95% fortalecerá a validade dos achados sobre sobrecarga cognitiva. Divergências sistemáticas entre métricas subjetivas e objetivas (e.g., baixa CL relatada mas alta atividade Téta frontal) serão interpretadas como indicadores de: (a) processos compensatórios não-conscientes; (b) variabilidade individual na metacognição do esforço; ou (c) desacoplamento temporal entre percepção e ativação neural. Esta hierarquização permite análises estatísticas focadas enquanto preserva a riqueza multidimensional da medição de CL.

*Resultado da Tarefa:*

* + - Taxa de sucesso na conclusão da tarefa de ajuste fino (por exemplo, convergência alcançada).
    - Métricas de qualidade do modelo ajustado (Acurácia, F1, ROUGE, Perplexidade) avaliadas em um conjunto de teste mantido constante.
    - Qualidade do resultado do fluxo de trabalho (avaliada usando métricas específicas da tarefa de análise de dados).

**3.4 Participantes**

Para a condução deste estudo, recrutou-se uma amostra intencional e diversificada de N=40 participantes, provenientes da relevante comunidade de Engenharia de Software da Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes (UnDF), um público diretamente alinhado aos objetivos da pesquisa. A amostra foi metodicamente estratificada com base no nível de expertise em LLMs, auto-relatado e subsequentemente validado através de critérios objetivos, resultando em dois grupos equilibrados: 20 Iniciantes e 20 Especialistas. Adicionalmente, envidou-se um esforço consciente para balancear a representação de gênero dentro de cada grupo de expertise, na medida das possibilidades de recrutamento, visando ampliar a validade externa dos achados.

* *Avaliação e Categorização de Expertise:* A expertise dos participantes em LLMs e QLoRA foi operacionalizada através de uma escala contínua composta (0-100), derivada da combinação ponderada de: (a) desempenho no 'Teste Estruturado de Conhecimento Conceitual e Resolução de Problemas em LLM/QLoRA', pré-pilotado e validado por dois avaliadores independentes (50% do escore); (b) anos de experiência prática autorrelatada (30%); e (c) portfólio de projetos relevantes concluídos (20%). A confiabilidade interavaliadores foi estabelecida com Kappa de Cohen = 0.89. Este teste, administrado online, compreendeu 8 questões de múltipla escolha avaliando conceitos fundamentais (e.g., função dos hiperparâmetros `r` e `lora\_alpha` em configurações QLoRA, princípios de quantização GGUF, arquiteturas de adaptação de baixo posto) e 3 cenários curtos de resolução de problemas abertos (e.g., identificar e explicar estratégias para diagnosticar erros de 'Out Of Memory' durante ajuste fino, interpretar logs de convergência problemática, propor otimizações de hiperparâmetros para modelos específicos).

## **Tabela 1** - Operacionalização da Escala Contínua de Expertise

| **Componente da Escala** | **Instrumento de Medida** | **Peso** | **Faixa Original** | **Normalização** | **Escala Final** |
| --- | --- | --- | --- | --- | --- |
| **Conhecimento Conceitual e Resolução de Problemas** | Teste Estruturado LLM/QLoRA (8 questões múltipla escolha + 3 cenários abertos) | 50% | 0-44 pontos | Linear: (pontuação/44) × 50 | 0-50 |
| **Experiência Prática Autorrelatada** | Questionário pré-sessão (anos de desenvolvimento/ajuste fino de LLMs) | 30% | 0-15+ anos | Logarítmica: min(log₂(anos+1)/4, 1) × 30 | 0-30 |
| **Portfólio de Projetos e Publicações** | Autorrelato validado (projetos IA/LLM + publicações revisadas por pares) | 20% | 0-10+ projetos/pubs | Sigmóide: (2/(1+e^(-x/2))-1) × 20 | 0-20 |
| **Escore Composto Final** | Soma ponderada dos componentes normalizados | 100% | - | Σ(componentes normalizados) | **0-100** |

### Critérios de Categorização Dicotômica

**Iniciantes**: Escore ≤ 50 (n=20)

**Especialistas**: Escore > 50 (n=20)

**Concordância com autorrelato inicial**: 95%

**Confiabilidade interavaliadores**: κ = 0.89

### Notas Metodológicas

**Normalização logarítmica** para experiência prática mitiga o efeito de outliers com experiência muito extensa

**Função sigmóide** para projetos/publicações evita saturação em casos de produtividade excepcional

**Validação cruzada** realizada através de correlação com performance em tarefa de diagnóstico pré-experimental (r = 0.73, p < 0.001)

* *Pontuação e Critérios de Avaliação:* O sistema de pontuação foi estruturado em escala de 4 pontos (0-4) por dois avaliadores independentes, com os seguintes critérios: Iniciantes (pontuação 0-1) demonstraram conhecimento limitado ou incorreto, como identificar apenas um hiperparâmetro corretamente sem explicação adequada ou sugerir soluções genéricas para OOM; Intermediários (pontuação 2) mostraram compreensão parcial, identificando 2-3 hiperparâmetros com explicações básicas corretas; Avançados (pontuação 3) demonstraram conhecimento sólido, explicando corretamente trade-offs entre hiperparâmetros; Especialistas (pontuação 4) exibiram compreensão profunda, articulando aspectos matemáticos complexos e técnicas avançadas. A confiabilidade interavaliadores foi alta (Kappa de Cohen = 0.89).
* *Categorização Final:* Os escores totais (0-44 pontos) contribuíram para uma medida contínua de expertise que foi posteriormente dicotomizada para análises específicas: participantes com pontuação ≤22 foram classificados como 'Iniciantes' enquanto aqueles com pontuação ≥23 foram categorizados como 'Especialistas', com concordância de 95% com o auto-relato inicial. Este pré-teste assegurou uma medida de base mais objetiva do entendimento prático e teórico relevante para as tarefas experimentais.
* *Considerações Estatísticas e Éticas:* A determinação do tamanho da amostra foi informada por uma análise de poder estatístico a priori conduzida com G\*Power (Faul et al., 2007). Para os Modelos Lineares Generalizados Mistos (GLMMs) planejados, que constituem a principal abordagem analítica para testar as hipóteses H1, H2 e H4, estimou-se o poder para detectar efeitos de interação e principais intra-sujeitos. Assumindo um nível alfa ($ \alpha ) de 0.05, poder desejado ($1-\beta$) de 0.80, uma correlação moderada ( \rho \approx 0.5 ) entre as medidas repetidas (consistente com a literatura em estudos de CL com desenhos intra-sujeito), e considerando o desenho fatorial misto (e.g., 2 níveis de Quantização x 3 níveis de Complexidade do Fluxo como fatores intra-sujeitos, e 2 níveis de Expertise como fator entre-sujeitos), a análise indicou que uma amostra de N=40 participantes seria suficiente para detectar tamanhos de efeito considerados médios a grandes ( \eta\_p^2 $ variando de aproximadamente 0.06 a 0.10) para os efeitos intra-sujeitos e interações de interesse. Especificamente, para um efeito de $ \eta\_p^2 = 0.06 $ (considerado um efeito médio segundo Cohen (1988)), o poder alcançado com N=40 é de aproximadamente 0.80. Este nível de sensibilidade foi considerado adequado para os objetivos do estudo, focando em efeitos de magnitude prática relevante no contexto da interação CL-CompL.

## **Tabela 3** - Parâmetros da Análise de Poder Amostral A Priori

| **Componente da Análise** | **Especificação** | **Justificativa/Referência** |
| --- | --- | --- |
| **Software e Método** |  |  |
| Ferramenta Estatística | G\*Power 3.1.9.7 (Faul et al., 2007) | Padrão-ouro para cálculos de poder em designs complexos |
| Família de Teste | F-tests: ANOVA (Repeated measures, within-between interaction) | Adequado para design fatorial misto 2×3×2 |
| **Parâmetros Estatísticos Core** |  |  |
| Nível Alfa (α) | 0.05 | Convenção científica padrão |
| Poder Alvo (1-β) | 0.80 | Recomendação Cohen (1988) para estudos comportamentais |
| **Estrutura do Design Experimental** |  |  |
| Fatores Between-Subjects | 1 (Expertise: 2 níveis) | Iniciante vs. Especialista |
| Fatores Within-Subjects | 2 (Quantização: 2 níveis; Fluxo: 3 níveis) | Q4 vs. Q8; Único vs. Sequencial vs. Paralelo |
| Total de Condições Intra-Sujeitos | 6 (2×3) | Todas as combinações Quantização×Fluxo |
| **Pressupostos de Correlação** |  |  |
| Correlação Entre Medidas Repetidas (ρ) | 0.50 | Baseado em estudos HCI/CL (Duran et al., 2022) |
| Faixa de Sensibilidade Testada | 0.30 - 0.70 | Análise robusta para diferentes cenários |
| **Tamanhos de Efeito e Detecção** |  |  |
| η²p Alvo (Efeito Médio) | 0.06 | Classificação Cohen: médio para estudos psicológicos |
| Cohen's f Equivalente | 0.253 | Calculado: f = √(η²p/(1-η²p)) |
| η²p Mínimo Detectável (N=40) | 0.045 | Para poder = 0.80, α = 0.05 |
| **Cenários de Poder Específicos** |  |  |
| Efeitos Principais (H1, H2) | Poder ≥ 0.82 | Para η²p = 0.06 |
| Interações Duplas (H4) | Poder ≥ 0.78 | Para η²p = 0.06 |
| Interação Tripla | Poder ≥ 0.65 | Para η²p = 0.06 (exploratória) |
| **Validação e Robustez** |  |  |
| Análise de Sensibilidade | Testada para ρ = 0.3, 0.5, 0.7 | Confirmou adequação para N=40 |
| Correção para Atrito | +15% buffer incluído | Previsão de 5-10% dropout |
| Validação Post-Hoc | Poder observado ≥ 0.85 | Para efeitos significativos detectados |

### **Notas Metodológicas Importantes:**

**Especificidade do Design**: O cálculo considerou a estrutura exata do experimento (design misto 2×3×2) ao invés de aproximações simplificadas.

**Literatura Base**: Os valores de correlação (ρ = 0.5) foram baseados em estudos similares de carga cognitiva em HCI (Duran et al., 2022; Paas et al., 2003).

**Análise Conservadora**: O tamanho de efeito alvo (η²p = 0.06) representa uma estimativa conservadora baseada em meta-análises de estudos de quantização (Nagel et al., 2021).  
**Validação Cruzada**: Análises de poder foram replicadas usando métodos analíticos (fórmulas de Cohen) e confirmadas via simulação Monte Carlo (1000 iterações).

O recrutamento e a condução do estudo seguiram protocolos aprovados pelo Comitê de Ética em Pesquisa institucional, com todos os participantes fornecendo consentimento informado por escrito antes das sessões experimentais de 60 minutos. Dados sensíveis foram anonimizados usando códigos de identificação, garantindo a privacidade dos participantes.

**3.5 Materiais e Aparatos**

Utilizou-se um ambiente de laboratório controlado com hardware e software padronizados para minimizar a variabilidade estranha e garantir a replicabilidade.

* **Hardware:** Conforme mencionado, duas estações de trabalho foram usadas como plataformas experimentais primárias, equipadas com:

GPU: NVIDIA GeForce RTX 3060 (12GB VRAM) em uma estação, NVIDIA GeForce RTX 4050 Laptop (6GB VRAM) na outra. A escolha destas GPUs visou representar hardware de prateleira, mas com capacidades distintas de VRAM, permitindo explorar o impacto das restrições de memória.

CPU, RAM, Armazenamento: Processadores e memória RAM adequados para suportar as tarefas (por exemplo, Intel Core i5/i7, 32GB+ RAM, SSD NVMe).

Sensores Neurofisiológicos:

* + - *Rastreador Ocular (POG):* **Tobii Eye Tracker 5**, um dispositivo baseado em tela, montado abaixo do monitor, capaz de rastrear movimentos oculares e diâmetro pupilar com alta frequência (por exemplo, 60-120 Hz).
    - *EEG:* **InteraXon Muse 2**, um headset EEG acessível com 4 eletrodos secos (posições aproximadas Fp1, Fp2, TP9, TP10) e um eletrodo de referência (Fpz). Embora tenha menor resolução espacial que sistemas de pesquisa, é não invasivo, fácil de configurar e demonstrou ser sensível a mudanças na carga cognitiva e no estado de atenção em estudos anteriores (Desai et al., 2022; Krigolson et al., 2017).
    - *PPG/SpO2:* **Samsung Galaxy Watch5**, um smartwatch comercial equipado com sensores PPG para medir a frequência cardíaca, a variabilidade da frequência cardíaca (HRV) e a saturação periférica de oxigênio (SpO2). A escolha desses sensores específicos foi baseada em um equilíbrio entre a qualidade dos dados fornecidos – validados na literatura para medição de correlatos de CL (Ma et al., 2024b; Yu et al., 2024; Belda-Lois, 2024; Traunmueller et al., 2024) – a facilidade de uso e configuração no ambiente experimental, o conforto do participante, a disponibilidade comercial e a compatibilidade com o framework de sincronização LSL.
* **Software:** A pilha de software foi padronizada em ambas estações experimentais:

*Sistema Operacional:* KDE Neon 6.0.

*Linguagem e Ambiente:* Python 3.10+.

*Plataforma CrossDebate:*

* + - Frontend: Aplicação React (v18+) com bibliotecas de UI (por exemplo, Material UI) e visualização (por exemplo, Plotly.js, Recharts).
    - Backend: FastAPI (v0.100+) executado com Uvicorn.

*Bibliotecas LLM/PEFT:* llama-cpp-python 0.1.77 (para inferência GGUF), transformers 4.39.3, peft 0.11.0, accelerate 0.27.2, bitsandbytes 0.39.0, torch 2.0.1 (para ajuste fino QLoRA), unsloth 2024.8 (para otimizações QLoRA).

*Bibliotecas de Processamento Fisiológico:* pylsl 1.16.2 (para streaming de dados LSL), mne-bids 0.14 (para pré-processamento e análise de EEG), neurokit2 0.2.9 (para processamento de PPG e outros sinais fisiológicos).

*Ferramentas de Monitoramento CompL:* psutil 6.0.0 (para CPU/RAM), scripts para consultar nvidia-smi 0.1.3 (para GPU VRAM, Utilização, Energia, Temperatura).

*Pacotes Estatísticos (para análise offline):* pandas 2.1.0, numpy 1.22.0, scipy 1.12.0, statsmodels 0.14.0, pingouin 0.5.2, scikit-learn 1.2.2, matplotlib 3.8.0. Versões específicas de todas as bibliotecas críticas foram fixadas (usando conda environment.yml) para garantir a reprodutibilidade do ambiente de software.

* **Modelos GGUF e Conjuntos de Dados:**

*Modelos:* Um conjunto de 60 modelos GGUF pré-quantizados foi preparado, compreendendo 30 modelos no formato Q4 e 30 no formato Q8. Estes modelos são variantes de arquiteturas abertas populares e eficientes, como Phi-4 e Mistral-7B, adequadas para execução local.

*Conjuntos de Dados:* Foram utilizados conjuntos de dados públicos e representativos para as tarefas de ajuste fino (por exemplo, datasets de código como MathQA-Python para geração de código) e para as tarefas de análise de dados dos fluxos multiagente. Os dados foram pré-processados e formatados adequadamente para o ajuste fino e a inferência.

A rigorosa calibração dos sensores, um passo indispensável para a fidedignidade dos dados, foi realizada no início de cada sessão experimental, seguindo estritamente as diretrizes metodológicas detalhadas por Blackmore et al. (2024), assegurando assim a padronização e a confiabilidade das medições subsequentes. Isso envolveu a calibração individualizada do rastreador ocular (utilizando o software Tobii Pro Lab) para garantir um mapeamento preciso do olhar e medições fidedignas da pupila, e a verificação meticulosa da qualidade do sinal do EEG (através da checagem de impedância dos eletrodos para assegurar boa condutividade e minimizar ruído) e PPG (confirmando o adequado contato com a pele para evitar artefatos de movimento). Adicionalmente, uma gravação de linha de base fisiológica em estado de repouso, com o participante relaxado e focado em um ponto neutro, foi efetuada durante alguns minutos para permitir a subsequente normalização dos dados, controlando assim a variabilidade interindividual intrínseca aos sinais fisiológicos. Conforme amplamente validado por Blackmore et al. (2024), esses procedimentos meticulosos são fundamentais para minimizar interferências externas e internas, assegurar a validade ecológica dos dados fisiológicos coletados em tempo real, especialmente em interações dinâmicas com sistemas de AI, e estabelecer referências individualizadas robustas que informam de maneira precisa as medições e interpretações subsequentes na plataforma CrossDebate.

**3.6 Procedimento Experimental**

O procedimento experimental foi padronizado para guiar os participantes através das diferentes condições experimentais, coletar dados de modo consistente e, fundamentalmente, garantir que a duração total da sessão para cada participante girasse em torno de 60 minutos.

* **Preparação e Consentimento (≈ 10 minutos):**
* Chegada do participante, recepção e breve explicação geral do estudo.
* Leitura e assinatura do Termo de Consentimento Livre e Esclarecido.
* Preenchimento de avaliação de expertise (auto-relato sobre experiência com LLMs).
* Colocação e ajuste dos sensores fisiológicos (headset EEG, smartwatch PPG/SpO2).
* Posicionamento do participante em frente à estação de trabalho e calibração do rastreador ocular (Tobii).
* Verificação inicial da qualidade dos sinais fisiológicos.
* Gravação de uma linha de base fisiológica curta (≈ 3-4 minutos) com o participante em repouso, olhando para uma tela neutra.

Antes do início dos blocos experimentais, os participantes foram questionados sobre fatores que poderiam influenciar as medidas fisiológicas, como consumo recente de cafeína e qualidade do sono na noite anterior. Essas informações foram registradas como covariáveis potenciais.

* **Treinamento e Familiarização (≈ 5 minutos):**
* Breve introdução à interface React da plataforma CrossDebate, destacando as principais áreas de interação (configuração, monitoramento, visualização).
* Visão geral concisa das tarefas a serem realizadas (ajuste fino de um modelo GGUF usando QLoRA, gerenciamento de um fluxo de trabalho multiagente).
* Explicação rápida dos conceitos chave (Quantização GGUF Q4/Q8, Complexidade do Fluxo).
* **Blocos Experimentais (≈ 40 minutos no total):**

Os participantes completaram um número definido de blocos de tarefas, correspondentes às condições experimentais intra-sujeitos (Nível de Quantização x Complexidade do Fluxo de Trabalho). A ordem dos blocos foi contrabalançada entre os participantes.

Dentro de cada bloco, o participante realizava uma tarefa específica definida pelo desenho experimental (como Bloco 1: Ajustar um modelo Q4 em uma tarefa de agente único; Bloco 2: Gerenciar um fluxo multiagente paralelo usando modelos Q8).

As tarefas envolveram interagir com a plataforma CrossDebate para: selecionar modelos GGUF do enxame, configurar parâmetros de ajuste fino QLoRA ou de fluxo de trabalho, iniciar a execução, monitorar o progresso (através de logs, métricas de CompL e visualizações de CL fornecidas pela plataforma), avaliar a qualidade das saídas e tomar decisões para refinar o processo.

Uma consideração metodológica refere-se à viabilidade de conduzir todas as manipulações experimentais (dois níveis de quantização GGUF × três níveis de complexidade do fluxo de trabalho, como fator intra-sujeitos) e, simultaneamente, coletar dados qualitativos ricos e significativos dentro de um período de aproximadamente 40 minutos alocado para os blocos experimentais. A densidade das condições experimentais em um tempo relativamente limitado pode, potencialmente, levar a interações mais superficiais por parte dos participantes em algumas das condições, o que poderia afetar a profundidade dos dados coletados e, consequentemente, a validade ecológica de algumas conclusões.

É fundamental enfatizar que o escopo experimental desta pesquisa foi deliberadamente limitado à investigação de interações elementares entre operadores humanos e sistemas LLM locais. Esta delimitação metodológica reconhece explicitamente as limitações temporais inerentes ao desenho experimental de 60 minutos por participante, necessárias para manter a viabilidade logística do estudo e garantir a qualidade das medições psicofisiológicas contínuas.

As tarefas implementadas na plataforma CrossDebate focaram estrategicamente em atividades básicas e fundamentais que constituem os blocos de construção de workflows mais complexos. Estas atividades foram cuidadosamente selecionadas por sua capacidade de serem realizadas e avaliadas adequadamente dentro do período experimental restrito, mantendo simultaneamente a relevância para contextos de aplicação mais amplos.

A operacionalização experimental incluiu cinco categorias principais de interações elementares. Primeiramente, a configuração inicial de parâmetros QLoRA básicos (taxa de aprendizado, rank, alpha) para um modelo pré-selecionado, representando aproximadamente 3-5 minutos de interação focada. Em segundo lugar, a avaliação qualitativa simples de saídas de modelos Q4 versus Q8 em prompts padronizados de classificação de sentimento, demandando cerca de 2-3 minutos por condição experimental.

A terceira categoria envolveu a execução de comandos básicos de orquestração multiagente através da interface React, como inicializar um pipeline sequencial de dois agentes ou configurar um fluxo paralelo simples envolvendo três agentes, requerendo 4-6 minutos por nível de complexidade. A quarta dimensão compreendeu a interpretação de métricas computacionais apresentadas em dashboards simplificados (uso de VRAM, temperatura da GPU, tokens por segundo), consumindo 1-2 minutos por avaliação sistemática.

Finalmente, a quinta categoria abrangeu ajustes reativos básicos em resposta a feedback da plataforma sobre estados de sobrecarga cognitiva ou computacional, como reduzir batch size ou simplificar prompts, representando intervenções focadas de 30-60 segundos cada.

Embora estas interações elementares não capturem a complexidade total de workflows profissionais de ajuste fino que podem se estender por horas ou dias em ambientes de produção, elas representam componentes fundamentais que são metodologicamente extrapoláveis para cenários mais complexos. Esta abordagem fornece insights valiosos sobre a dinâmica básica CL-CompL que constitui a base teórica necessária para informar o design de sistemas mais sofisticados e de maior escala.

A escolha estratégica por focar em interações elementares proporcionou múltiplas vantagens metodológicas significativas. Permitiu maior controle experimental sobre variáveis confundidoras, assegurou replicabilidade dos procedimentos através de diferentes participantes e condições, e facilitou a coleta de dados psicofisiológicos de alta qualidade em ambiente controlado.

Simultaneamente, esta abordagem manteve a validade conceitual fundamental para o desenvolvimento de diretrizes de design ergonômico para sistemas de AI locais, estabelecendo princípios básicos que podem ser escalados e adaptados para aplicações mais complexas em contextos profissionais e educacionais diversos.

Durante toda a duração dos blocos experimentais, os dados fisiológicos (POG, EEG, PPG, SpO2), as métricas de CompL (VRAM, GPU util, energia, temp) e os registros de interação com a interface foram registrados continuamente e sincronizados via LSL.

Em segmentos pré-determinados e curtos (para não exceder o tempo), os participantes foram solicitados a pensar em voz alta.

* **Avaliações Pós-Bloco (≈ 5 minutos no total):**

Imediatamente após a conclusão de cada bloco de tarefas, os participantes completaram o questionário NASA-TLX Adaptado e a SEQ, referentes à(s) condição(ões) que acabaram de experienciar. Isso foi feito diretamente na plataforma. O tempo foi estritamente gerenciado para esta fase.

* **Finalização e Debriefing (Tempo Mínimo/Pós-Sessão):**
* Remoção dos sensores.
* Agradecimento ao participante pela sua contribuição.
* Oportunidade para perguntas finais. Um debriefing mais detalhado sobre os objetivos específicos e hipóteses é fornecido por escrito após a sessão para economizar tempo.

Este procedimento padronizado visou maximizar a coleta de dados relevantes dentro da restrição de 60 minutos, mantendo o engajamento do participante e minimizando a fadiga, ao mesmo tempo que permitia a manipulação controlada das variáveis independentes e a medição confiável das variáveis dependentes.

**3.7 Medição de Dados**

Conforme detalhado na Seção 3.3 (Variáveis Dependentes), múltiplos fluxos de dados foram coletados para operacionalizar CL, CompL e resultados da tarefa. A sincronização temporal exata via LSL foi um aspecto metodológico fundamental. Cada fluxo de dados (EEG, PPG, POG, eventos da plataforma, registros de interação, métricas de CompL) recebeu timestamps LSL comuns, permitindo o alinhamento exato de eventos cognitivos (detectados fisiologicamente ou subjetivamente) com eventos da plataforma (início/fim de épocas de ajuste fino ou erros de CompL) e ações do usuário (interações na interface React).

Isso permitiu análises por fase da tarefa. Por exemplo, foi possível calcular a CL fisiológica média e o pico de VRAM especificamente durante a fase de "avaliação de saída do modelo Q4" versus a fase de "configuração de parâmetros QLoRA", permitindo investigar como diferentes subtarefas impactaram o conjunto CL-CompL.

O pré-processamento dos sinais fisiológicos brutos foi essencial para extrair destacáveis características e remover artefatos. Pipelines de processamento padronizados foram implementados usando mne-bids 0.14 para EEG (filtragem, rejeição de segmentos ruidosos, cálculo de potência espectral) e neurokit2 0.2.9 para PPG (detecção de picos R, cálculo de métricas HRV no domínio do tempo e da frequência, filtragem de SpO2). Da mesma forma, os dados brutos de rastreamento ocular foram processados para extrair métricas de fixação, sacada e diâmetro pupilar, usando a biblioteca pygaze 0.9.1. Os dados de CompL foram registrados pelo backend FastAPI em intervalos regulares de 5 segundos e agregados (média, pico) por fase da tarefa.

**3.8 Rigor e Validade**

Foram tomadas medidas proativas ao longo do planejamento e execução do estudo para maximizar o rigor metodológico e abordar potenciais ameaças à validade dos resultados.

**Validade Interna:** Refere-se à confiança de que as relações observadas entre VIs e VDs são causais. Foi fortalecida através de: (1) Controle Experimental: Padronização de tarefas, hardware, software e procedimentos; (2) Manipulação Sistemática: VIs (Quantização, Fluxo) foram diretamente manipuladas pelo pesquisador; (3) Contrabalanceamento: Mitigou efeitos de ordem para os fatores intra-sujeitos; (4) Medição de Covariáveis: Medição da expertise como fator entre-sujeitos para controlar seu efeito.

**Validade de Construto:** Refere-se à adequação da operacionalização das variáveis teóricas (CL, CompL, Expertise). Foi fortalecida através de: (1) Triangulação: Uso de múltiplas medidas (subjetivas, fisiológicas, comportamentais) para CL; (2) Instrumentos Estabelecidos: Uso do NASA-TLX Adaptado e de correlatos fisiológicos de CL bem documentados na literatura; (3) Métricas Diretas: Uso de métricas de sistema padrão e diretas para CompL (VRAM, Utilização, Energia); (4) Operacionalização Clara: Definição precisa dos níveis de Expertise e Complexidade do Fluxo.

**Validade Externa (Generalização):** Refere-se à aplicabilidade dos achados a outros contextos, populações e tempos. Foi considerada através de: (1) Representatividade: Escolha de tarefas (ajuste fino QLoRA, gerenciamento de fluxo multiagente), modelos (GGUF Q4/Q8 de arquiteturas populares), hardware (GPUs de prateleira) e participantes (incluindo iniciantes e especialistas) que refletem cenários realistas de uso de AI local; (2) Limitações Reconhecidas: A validade externa dos achados apresenta delimitações específicas que devem ser consideradas na aplicação dos resultados. Os achados são primariamente aplicáveis a: (a) configurações de hardware similares às testadas (GPUs NVIDIA série RTX com 6-12GB VRAM); (b) formatos de quantização GGUF Q4/Q8; (c) técnicas de ajuste fino QLoRA; e (d) populações com características demográficas e de expertise similares à amostra (estudantes e profissionais de Engenharia de Software). A generalização para outras arquiteturas de GPU, diferentes runtimes LLM, ou populações com perfis técnicos distintos requer validação empírica adicional, pois os padrões de interdependência CL-CompL podem variar significativamente em função de diferenças nos gargalos computacionais e na expertise do operador. Os resultados são específicos para os modelos, tarefas e hardware testados, embora os princípios subjacentes da interação CL-CompL devam ser mais gerais.

**Confiabilidade:** Refere-se à consistência e replicabilidade das medidas e procedimentos. Foi assegurada através de: (1) Padronização: Procedimentos experimentais e instruções idênticas para todos os participantes (dentro de cada condição); (2) Registro Automatizado: Coleta de dados fisiológicos, de CompL e de interação realizada automaticamente pela plataforma; (3) Processamento Padronizado: Uso de pipelines de análise de dados e processamento de sinais bem definidos e documentados (usando bibliotecas padrão); (4) Análise Replicável: Planejamento de análises estatísticas usando métodos padrão (GLMMs).

Apesar dessas medidas, limitações inerentes foram reconhecidas. A interpretação de sinais fisiológicos como indicadores de CL é complexa e pode ser influenciada por fatores individuais e contextuais não controlados, incluindo variações circadianas, diferenças metabólicas individuais, sensibilidade diferencial a estressores cognitivos e histórico de exposição a tecnologias de AI. O ambiente de laboratório, embora controlado, difere de um ambiente de trabalho real, onde distrações externas, pressões organizacionais, multitarefa e familiaridade com ferramentas específicas podem modular significativamente a interação CL-CompL. A amostra, embora diversificada em termos de expertise, foi limitada em tamanho e restrita a uma população universitária específica, limitando a generalização para outros contextos profissionais, faixas etárias e culturas organizacionais.

A duração padronizada de 60 minutos por sessão experimental foi deliberadamente selecionada para otimizar a qualidade e confiabilidade das medições psicofisiológicas crossmodais, minimizando artefatos de degradação de sinal (como deslocamento de eletrodos EEG ou interferência por sudorese) e controlando efeitos confundidores de fadiga acumulativa que poderiam comprometer a interpretação de correlatos neurais específicos da CL. Este protocolo temporal captura efetivamente adaptações táticas de curto prazo (como alternância entre estratégias cognitivas durante transições Q4/Q8) mas não abrange adaptações estratégicas ou baseadas em desenvolvimento de expertise que emergem através de exposição prolongada a sistemas de AI. Consequentemente, nossos achados são diretamente aplicáveis a cenários de interação inicial e episódica com sistemas GGUF, mas a generalização para padrões de uso sustentado em ambientes profissionais requer validação longitudinal complementar que investigue como processos adaptativos de diferentes escalas temporais moderam a interdependência CL-CompL observada.

Esta metodologia detalhada, de métodos mistos, foi projetada para fornecer uma investigação válida, confiável e matizada da complexa interação CL-CompL no domínio prático e relevante do gerenciamento local de LLMs GGUF através da plataforma CrossDebate, dentro de uma restrição temporal realista de 60 minutos por participante.

**4. ANÁLISE DE DADOS**

Esta seção descreve o plano detalhado e sistemático para analisar os dados crossmodais coletados através da metodologia de métodos mistos descrita na Seção 3. O objetivo central da análise foi abordar as questões de pesquisa e testar formalmente as hipóteses sobre a interdependência dinâmica entre a Carga Cognitiva (CL) humana e a Carga Computacional (CompL) da plataforma. Este processo ocorreu no contexto do ajuste fino local (QLoRA) e do gerenciamento de fluxos de trabalho multiagente envolvendo o enxame de 60 modelos LLM, orquestrados pela plataforma CrossDebate em hardware de prateleira. O plano analítico abrangeu desde a preparação e limpeza dos dados sincronizados via LSL até análises descritivas, inferenciais (utilizando principalmente Modelos Lineares Generalizados Mistos - GLMMs para lidar com a estrutura aninhada dos dados), análises de correlação e regressão (incluindo modelagem não linear para investigar H3), análise qualitativa dos protocolos de pensamento em voz alta (via Análise Temática) e, primordialmente, a integração sistemática dos achados quantitativos e qualitativos para uma interpretação holística e rica do fenômeno CL-CompL.

A complexidade metodológica inerente à investigação da interdependência CL-CompL requer uma abordagem analítica que transcenda métodos estatísticos convencionais, incorporando princípios da teoria de sistemas dinâmicos e da análise de séries temporais multivariadas. A natureza temporal dos dados coletados - onde estados cognitivos e computacionais evoluem continuamente durante as sessões de 60 minutos - demanda técnicas analíticas que capturem não apenas relações estáticas entre variáveis, mas também padrões dinâmicos de retroalimentação e adaptação. Esta perspectiva sistêmica reconhece que a CL e CompL não são meramente variáveis correlacionadas, mas componentes de um sistema acoplado onde mudanças em um domínio propagam efeitos cascata no outro, criando loops de retroalimentação que podem resultar em estados de equilíbrio dinâmico ou, alternativamente, em oscilações ou instabilidades comportamentais.

A validação da robustez analítica constitui um pilar fundamental desta investigação, especialmente considerando a natureza pioneira da pesquisa sobre interdependência CL-CompL em sistemas de AI locais. A estratégia de validação adota uma abordagem multifacetada que inclui validação cruzada temporal (dividindo os dados em períodos de treinamento e teste), bootstrap paramétrico para estimar intervalos de confiança robustos, e análise de sensibilidade sistemática para avaliar como variações nos parâmetros dos modelos afetam as conclusões principais. Adicionalmente, a implementação de técnicas de permutação para testar a significância estatística sob hipóteses nulas específicas garante que os padrões observados não sejam artefatos de dependências espúrias nos dados. Esta robustez metodológica é primordial para estabelecer a validade externa dos achados e sua aplicabilidade a contextos mais amplos de interação humano-máquina em sistemas de AI.

**4.1 Visão Geral da Estratégia Analítica**

A análise seguiu um desenho convergente de métodos mistos (Creswell e Clark, 2017; Greene et al., 1989). Neste desenho, os dados quantitativos (fisiológicos: POG, EEG, PPG, SpO2; subjetivos: NASA-TLX Adaptado, SEQ; comportamentais: registros de interação; CompL: VRAM, GPU util, energia, temperatura; desempenho da tarefa) e qualitativos (transcrições dos protocolos de pensamento em voz alta) foram coletados simultaneamente durante as sessões experimentais de 60 minutos. Subsequentemente, foram analisados separadamente. Finalmente, os resultados das análises quantitativa e qualitativa foram formalmente mesclados, comparados e integrados durante a fase de interpretação.

A adoção de um framework de análise bayesiana como complemento às abordagens frequentistas tradicionais oferece vantagens significativas para a investigação da interdependência CL-CompL, particularmente na incorporação de conhecimento prévio sobre limitações cognitivas humanas e restrições computacionais. A estatística bayesiana permite a especificação de priors informativos baseados na literatura sobre Teoria da Carga Cognitiva e benchmarks de performance de hardware, resultando em estimativas posteriores que balanceiam evidência empírica com conhecimento teórico estabelecido. Esta abordagem é especialmente valiosa na modelagem de relações não-lineares entre CL e CompL, onde priors bem especificados podem regularizar modelos complexos e prevenir overfitting, enquanto proporcionam estimativas de incerteza mais realistas que refletem tanto variabilidade estatística quanto incerteza epistêmica sobre os mecanismos subjacentes.

A estratégia de triangulação metodológica implementada nesta análise vai além da simples combinação de métodos quantitativos e qualitativos, constituindo uma abordagem epistemológica que reconhece a natureza multidimensional da experiência humana na interação com sistemas de AI. A triangulação opera em múltiplos níveis: triangulação de dados (fisiológicos, comportamentais, subjetivos), triangulação de métodos (estatísticos, qualitativos, computacionais), triangulação de pesquisadores (múltiplos analistas independentes), e triangulação teórica (perspectivas da psicologia cognitiva, ciência da computação e ergonomia). Esta abordagem multifacetada não apenas aumenta a validade interna dos achados, mas também facilita a identificação de padrões emergentes que poderiam permanecer ocultos em análises unidimensionais, contribuindo para uma compreensão mais holística dos mecanismos que governam a interdependência CL-CompL.

A implementação de técnicas de machine learning interpretável como complemento às análises estatísticas tradicionais oferece insights únicos sobre padrões complexos na interação CL-CompL que podem não ser capturados por modelos lineares ou lineares generalizados. Algoritmos como Random Forest com importância de variáveis permutada, SHAP (SHapley Additive exPlanations) values para modelos de ensemble, e modelos aditivos generalizados com termos de interação fornecem uma perspectiva complementar sobre quais variáveis e combinações de variáveis são mais preditivas de estados cognitivos e computacionais específicos. Esta abordagem de machine learning interpretável é particularmente valiosa para identificar interações de ordem superior entre múltiplas modalidades de dados que poderiam ser perdidas em análises univariadas ou bivariadas tradicionais, contribuindo para o desenvolvimento de modelos preditivos da dinâmica CL-CompL que possam informar design futuro de sistemas de AI mais ergonômicos.

Esta abordagem de triangulação permitiu que os pontos fortes de um tipo de dado compensassem as fraquezas do outro, levando a uma compreensão mais profunda e matizada da interação CL-CompL. Por exemplo, um importante aumento estatisticamente em um indicador fisiológico de CL (quantitativo) foi explicado por temas emergentes da análise qualitativa que descreveram dificuldades específicas de interpretação ou frustração com a plataforma. A estratégia analítica compreendeu as seguintes etapas principais:

**1) Preparação e Limpeza de Dados:** Transformação dos dados brutos crossmodais e sincronizados via LSL em um formato estruturado, limpo e confiável, abordando questões como dados ausentes, outliers, sincronização temporal e agregação de dados fisiológicos e de CompL em épocas significativas.

**2) Análise Quantitativa Descritiva:** Sumarização das características centrais dos dados de CL, CompL e desempenho através de estatísticas descritivas (médias, medianas, desvios padrão, etc.) e visualizações (histogramas, gráficos de violino, gráficos de barras) para obter uma visão geral inicial dos padrões e distribuições, estratificados pelas variáveis independentes.

**3) Análise Quantitativa Inferencial (Teste de Hipóteses):** Utilização de Modelos Lineares Generalizados Mistos (GLMMs) como principal ferramenta estatística para testar formalmente as hipóteses H1 (efeito da Quantização), H2 (efeito da Complexidade do Fluxo) e H4 (moderação pela Expertise) sobre as variáveis dependentes (indicadores de CL, CompL, resultados da tarefa), controlando adequadamente a estrutura hierárquica e de medidas repetidas dos dados.

**4) Análise de Correlação e Regressão (Investigação da Interdependência):** Exploração da natureza da relação entre múltiplos indicadores de CL e CompL (abordando RQ3) e teste formal da hipótese de não linearidade (H3), utilizando correlações e modelos de regressão (incluindo termos polinomiais e Modelos Aditivos Generalizados Mistos - GAMMs).

**5) Análise Qualitativa (Análise Temática):** Aplicação da Análise Temática (Braun e Clarke, 2006) aos dados textuais (transcrições dos protocolos de pensamento em voz alta) para identificar, analisar e relatar padrões (temas) recorrentes relacionados às experiências subjetivas, estratégias cognitivas, desafios percebidos e interpretações dos participantes sobre a interação CL-CompL e o uso da plataforma CrossDebate.

**6) Integração de Métodos Mistos:** Síntese sistemática dos achados quantitativos e qualitativos na fase de discussão para fornecer uma interpretação contextualizada. Isso envolveu usar os resultados de um método para explicar, expandir ou refinar os resultados do outro, buscando convergência, complementaridade ou explicando dissonâncias.

A análise de dados longitudinais de alta frequência, característica dos dados psicofisiológicos coletados, requer consideração cuidadosa de questões de autocorrelação temporal e heteroscedasticidade que podem violar pressupostos de modelos estatísticos padrão. A implementação de modelos de séries temporais multivariadas, incluindo modelos de vetores autorregressivos (VAR) e modelos de espaço de estados, permite capturar tanto dependências temporais quanto relações dinâmicas entre CL e CompL. Estas abordagens são complementadas por análises de coerência espectral que identificam bandas de frequência onde CL e CompL exibem sincronização, revelando ritmos intrínsecos da interação humano-máquina que operam em diferentes escalas temporais - desde flutuações de alta frequência relacionadas a microtarefas específicas até padrões de baixa frequência associados a estados cognitivos sustentados ao longo das sessões experimentais.

**4.2 Preparação e Limpeza de Dados**

Esta fase inicial e criticamente fundamental foi essencial para garantir a qualidade superior, a integridade inquestionável e a adequação completa dos dados brutos. Tal procedimento preparou-os meticulosamente para as análises subsequentes, assegurando assim a confiabilidade e a validade dos achados da pesquisa, seguindo rigorosamente as melhores práticas estabelecidas na literatura científica para o pré-processamento de dados crossmodais e de séries temporais (Hair et al., 2010; Pyle, 1999; Keogh e Kasetty, 2003).

A implementação de um pipeline de controle de qualidade de dados baseado em princípios da teoria da informação garante que apenas dados com alta fidelidade informacional sejam incluídos nas análises subsequentes. Este pipeline incorpora métricas de entropia para avaliar a riqueza informacional de sinais fisiológicos, detectores automáticos de artefatos baseados em máquinas de vetores de suporte (SVM) treinadas em dados rotulados por especialistas, e algoritmos de detecção de outliers multivariados que consideram tanto magnitude quanto padrões temporais anômalos. A validação cruzada deste pipeline com dados de referência estabelece limites de confiança estatística que permitem rejeição automática de segmentos de dados comprometidos, garantindo que as análises subsequentes operem com dados de qualidade.

**1) Agregação, Estruturação e Sincronização:** Dados de fontes díspares (arquivos LSL contendo EEG, PPG, POG; logs do backend FastAPI registrando métricas de CompL e eventos da plataforma; logs do frontend React registrando interações do usuário; arquivos de questionários pós-bloco; transcrições de áudio) foram reunidos. Um conjunto de dados de análise primário foi criado (usando a biblioteca pandas 2.1.0), onde cada linha representou uma relevante unidade de análise (como uma fase específica como "configuração QLoRA"). As colunas continham identificadores (ParticipanteID, BlocoID, ÉpocaID), as variáveis independentes (Nível de Quantização, Complexidade do Fluxo, Expertise), as variáveis dependentes processadas e covariáveis relevantes (como linha de base fisiológica). O alinhamento temporal preciso, garantido pelos timestamps LSL, foi relevante para vincular corretamente os dados fisiológicos, de CompL e de interação a eventos específicos da tarefa.

**2) Processamento de Sinais Fisiológicos:** Os dados fisiológicos brutos exigiram pré-processamento substancial para remover artefatos e extrair importantes características, usando pipelines padronizados:

* EEG (via *mne-bids 0.14*): Filtragem passa-banda (por exemplo, 1-40 Hz), remoção de artefatos (por exemplo, piscadas, movimentos musculares) usando métodos como Análise de Componentes Independentes (ICA), segmentação em épocas (5 segundos), cálculo da Densidade Espectral de Potência (PSD) usando métodos como o de Welch, e extração da potência média (absoluta ou relativa) nas bandas de interesse (Téta, Alfa, Beta, Gama) para os canais relevantes (frontal, parietal).
* *PPG (via neurokit2 0.2.9):* Filtragem do sinal PPG, detecção dos picos de pulso, cálculo dos intervalos inter-batimentos (IBI), interpolação e reamostragem para obter um sinal equidistante, e cálculo de métricas HRV no domínio do tempo (RMSSD, SDNN) e da frequência (potência LF, HF, razão LF/HF) em janelas deslizantes (por exemplo, 120 segundos com sobreposição).
* *POG (via software da Tobii):* Filtragem e interpolação de dados de diâmetro pupilar, cálculo do diâmetro médio por época (possivelmente normalizado pela linha de base individual), detecção de fixações e sacadas usando algoritmos baseados em velocidade ou dispersão, e cálculo de métricas como duração/frequência de fixação (total e por AOI) e frequência/amplitude de sacada.
* *SpO2:* Filtragem e cálculo da média por época.

**3) Segmentação e Agregação:** Os dados contínuos (fisiológicos processados, CompL) foram segmentados em épocas alinhadas a eventos específicos da tarefa (marcados via LSL durante o experimento, por exemplo, início/fim de uma época de ajuste fino, apresentação de uma saída de modelo para avaliação). Características estatísticas (média, mediana, desvio padrão, mínimo, máximo) foram calculadas para cada VD dentro de cada segmento/época para reduzir a dimensionalidade e criar as variáveis para o conjunto de dados de análise. A escolha da duração do segmento foi baseada na dinâmica esperada do sinal e da tarefa (por exemplo, HRV requer janelas mais longas que a pupila, como 120 segundos).

**4) Tratamento de Dados Ausentes:** A extensão e o padrão de dados ausentes foram avaliados (como artefatos irrecuperáveis). Se a quantidade de dados ausentes fosse substancial (>5-10%) e o padrão sugerisse que não eram Completamente Aleatórios (MCAR), a Imputação Múltipla por Equações Encadeadas (MICE) foi considerada como a estratégia primária. Usando o IterativeImputer do scikit-learn 1.2.2, múltiplos conjuntos de dados imputados (por exemplo, m=20-50) foram criados. As análises estatísticas foram realizadas em cada conjunto imputado e os resultados (coeficientes, erros padrão) foram agrupados usando as regras de Rubin (Rubin, 1987) para obter estimativas finais e intervalos de confiança que refletiram a incerteza da imputação. Se os dados ausentes foram mínimos, a exclusão listwise foi considerada, mas com cautela.

**5) Detecção e Tratamento de Outliers:** Outliers univariados (valores extremos em uma única VD) e multivariados (combinações incomuns de valores em múltiplas VDs) foram examinados usando métodos gráficos (violinos, KDE) e estatísticos (distância de Mahalanobis). A principal abordagem foi usar métodos estatísticos robustos (como GLMMs, que são menos sensíveis a outliers do que modelos OLS simples). A técnica da Winsorização (substituir valores extremos por valores no limite de um percentil, por exemplo, 95º) foi aplicada (Barnett e Lewis, 1994; Osborne e Overbay, 2004). A remoção completa de outliers foi evitada sempre que possível, pois eles continham informações valiosas sobre a variabilidade do fenômeno.

**6) Verificação de Pressupostos e Transformação:** As variáveis dependentes contínuas foram verificadas quanto aos pressupostos dos modelos estatísticos planejados (principalmente normalidade e homocedasticidade dos resíduos para GLMMs com distribuição Gaussiana). Histogramas, gráficos Q-Q e testes formais (Shapiro-Wilk, Levene) foram usados. Quando os pressupostos foram violados, transformações de dados (por exemplo, logarítmica) foram aplicadas à VD para tentar estabilizar a variância. Quando as transformações não foram eficazes, GLMMs com distribuições não-Gaussianas (Gama para dados positivos assimétricos, Binomial para proporções) foram considerados. Preditores contínuos nos modelos de regressão (por exemplo, métricas de CompL ao prever CL) foram centrados na média para melhorar a interpretabilidade dos interceptos e coeficientes de interação e reduzir a multicolinearidade (Aiken et al., 1991).

**4.3 Estatísticas Descritivas**

Antes de realizar análises inferenciais, foram calculadas e examinadas estatísticas descritivas para todas as variáveis chave (VDs de CL, CompL, desempenho; VIs). Isso incluiu medidas de tendência central (média, mediana), dispersão (desvio padrão, intervalo interquartil - IIQ, mínimo, máximo) e distribuição (frequências, porcentagens para variáveis categóricas).

Essas estatísticas foram calculadas tanto para a amostra geral quanto estratificadas pelos níveis das variáveis independentes (por exemplo, CL média para Q4 vs Q8; pico de VRAM para Iniciantes vs Especialistas em fluxos paralelos). Os resultados foram visualizados usando os seguintes tipos de gráficos:

* **Histogramas:** Para visualizar a distribuição de VDs contínuas (e verificar normalidade).
* **Gráficos Violino:** Para comparar a distribuição (mediana, quartis, outliers) de VDs contínuas entre os diferentes níveis das VIs categóricas (Quantização, Fluxo, Expertise).
* **Gráficos de Barras (com Barras de Erro/ICs):** Para comparar médias de VDs contínuas entre grupos ou para exibir frequências/porcentagens de VDs categóricas.
* **Gráficos Q-Q (Quantil-Quantil):** Para uma avaliação gráfica mais detalhada da conformidade dos dados de VDs contínuas com distribuições teóricas, notadamente a distribuição normal, sendo uma ferramenta chave na verificação de normalidade.
* **Gráficos de Densidade de Kernel (KDE):** Utilizados tanto para visualizar a distribuição de VDs contínuas individuais (complementando os histogramas na verificação de normalidade e oferecendo uma visão mais suave da forma da distribuição) quanto para explorar visualmente a relação entre pares de variáveis contínuas (por exemplo, entre um indicador de CL e um de CompL). No caso bivariado, estimaram a densidade de probabilidade conjunta, o que permitiu identificar regiões onde os dados eram mais concentrados e visualizar a forma da distribuição bivariada, incluindo potenciais agrupamentos ou tendências não lineares.

Essa análise descritiva inicial foi primordial para: (a) obter uma compreensão intuitiva dos dados, (b) identificar relações potenciais, (c) verificar a eficácia das manipulações experimentais (por exemplo, confirmar que Q4 realmente usou menos VRAM que Q8), (d) detectar possíveis anomalias nos dados, e (e) informar a especificação dos modelos estatísticos inferenciais subsequentes.

**4.4 Estatísticas Inferenciais (Teste de Hipóteses via GLMMs)**

Modelos Lineares Generalizados Mistos (GLMMs) foram a principal ferramenta analítica para testar formalmente as hipóteses H1, H2 e H4. Os GLMMs, uma extensão flexível dos modelos lineares generalizados (GLMs) que acomodam estruturas de dependência como medidas repetidas e agrupamento hierárquico (Baayen et al., 2008; Bolker et al., 2009), foram ajustados usando pacotes como statsmodels 0.14.0. A estimação dos parâmetros foi realizada utilizando Máxima Verossimilhança Restrita (REML) para os componentes de variância e Máxima Verossimilhança (ML) para comparações de modelos via Testes de Razão de Verossimilhança (LRTs), adotando um critério objetivo de rejeição do modelo mais simples para p < 0.05 e critérios de convergência padrão (por exemplo, tolerância de 1e-6 para a função de log-verossimilhança). A significância dos efeitos fixos foi avaliada comparando o modelo completo com modelos reduzidos sem o termo de interesse via LRTs. Após o ajuste de cada GLMM, os pressupostos do modelo (normalidade e homocedasticidade dos resíduos para modelos Gaussianos) foram verificados visualmente (gráficos Q-Q, resíduos vs. ajustados) e com testes formais (como Shapiro-Wilk ou Levene). Violações significativas (p < 0.05) levaram à aplicação de transformações como Box-Cox (para normalização) e logarítmica (para estabilização de variância) ou à reconsideração da família de distribuição e função de ligação do modelo, conforme práticas metodológicas robustas (Dettmers et al., 2023).

**Especificação do Modelo:** Para cada VD quantitativa chave (por exemplo, escore total ponderado do NASA-TLX, diâmetro pupilar médio por época, potência Téta frontal, RMSSD, pico de VRAM, tempo na subtarefa, acurácia do modelo ajustado), um GLMM separado foi ajustado usando pacotes estatísticos como statsmodels 0.14.0. A especificação do modelo incluiu:

● *Distribuição e Função de Ligação:* Escolhidas com base na natureza da VD (por exemplo, Gaussiana com ligação identidade para VDs contínuas aproximadamente normais; Binomial com ligação logit para VDs binárias como sucesso/falha; Gama com ligação log para VDs contínuas positivas e assimétricas como tempos de reação).

● *Efeitos Fixos:* Termos representando as VIs principais (Nível de Quantização - fator com 2 níveis, Complexidade do Fluxo - fator com 3 níveis, Expertise - fator com 2 níveis), todas as interações de duas vias (Quantização\*Fluxo, Quantização\*Expertise, Fluxo\*Expertise) e a interação de três vias (Quantização\*Fluxo\*Expertise). Covariáveis relevantes (por exemplo, linha de base fisiológica pré-tarefa, ordem do bloco) também foram incluídas como efeitos fixos para controlar sua influência. Fatores categóricos foram codificados usando esquemas de contraste como a codificação de tratamento para comparar cada nível com uma linha de base.

● *Efeitos Aleatórios:* Para contabilizar a não independência das observações dentro dos participantes e a variabilidade basal entre eles, pelo menos um intercepto aleatório para ParticipanteID foi incluído em todos os modelos. Além disso, inclinações aleatórias para os fatores intra-sujeitos (Quantização, Fluxo) por ParticipanteID foram consideradas. A inclusão de inclinações aleatórias permitiu que o efeito de uma VI variasse entre os participantes. A estrutura de efeitos aleatórios "maximal" (incluindo todas as inclinações aleatórias justificadas pelo desenho) foi inicialmente considerada, mas simplificada se o modelo falhasse em convergir (avaliado via Testes de Razão de Verossimilhança - LRTs - comparando modelos aninhados).

Dada a natureza exploratória deste estudo e o conjunto abrangente de variáveis dependentes (VDs) coletadas para operacionalizar múltiplas facetas da Carga Cognitiva (CL), Carga Computacional (CompL) e desempenho da tarefa (detalhadas na Seção 3.3), implementou-se uma estratégia hierárquica de correção para comparações múltiplas que equilibra rigor estatístico com preservação da sensibilidade analítica.

Dada a natureza exploratória deste estudo e o conjunto abrangente de variáveis dependentes coletadas, implementou-se uma estratégia hierárquica de correção para comparações múltiplas. Para as análises primárias testando as hipóteses centrais (H1-H4), aplicou-se o procedimento de Benjamini-Hochberg com α = 0.05 para controlar a Taxa de Falsas Descobertas (FDR) especificamente nos testes de: (a) efeitos principais da Quantização e Complexidade do Fluxo nas métricas primárias de CL e CompL; e (b) termos de interação envolvendo Expertise como moderador. Esta abordagem preserva o rigor estatístico para as conclusões principais enquanto mantém sensibilidade analítica para descobertas emergentes. Esta correção foi aplicada especificamente aos valores-p derivados dos efeitos principais e interações teoricamente fundamentados: (a) efeito principal da Quantização GGUF nas métricas primárias de CL (Escore Total Ponderado NASA-TLX, potência da banda Téta frontal) e CompL (pico de uso VRAM, consumo energético); (b) efeito principal da Complexidade do Fluxo nas mesmas métricas; e (c) termos de interação envolvendo Expertise como moderador (Quantização×Expertise, Fluxo×Expertise).

Para comparações post-hoc subsequentes a efeitos principais ou interações significativas dentro de modelos GLMM individuais (e.g., contrastes pareados entre os três níveis de 'Complexidade do Fluxo'), empregou-se o método HSD de Tukey para controlar a Taxa de Erro Familiar (FWER) dentro de cada família específica de comparações, conforme práticas metodológicas estabelecidas para análises de seguimento em delineamentos fatoriais (Bretz et al., 2011).

As análises exploratórias de métricas secundárias e investigações de relações CL-CompL (RQ3) foram conduzidas com o nível α nominal de 0.05, mas interpretadas com cautela adicional dada a natureza confirmatória limitada destes testes. Esta abordagem em camadas assegurou proteção estatística rigorosa para as conclusões principais enquanto preservou a capacidade de detectar padrões emergentes relevantes para futuras investigações da interdependência CL-CompL.

**Teste de Hipóteses:** As hipóteses foram testadas examinando a significância estatística dos efeitos fixos no modelo GLMM ajustado:

●*H1 (Efeito da Quantização):* Testada pela significância do efeito principal do fator Nível de Quantização.

●*H2 (Efeito da Complexidade do Fluxo):* Testada pela significância do efeito principal do fator Complexidade do Fluxo.

●*H4 (Moderação pela Expertise):* Testada pela significância dos termos de interação envolvendo o fator Expertise (Quantização\*Expertise, Fluxo\*Expertise, e a interação de três vias). A significância foi avaliada usando LRTs (comparando o modelo completo com um modelo reduzido sem o termo de interesse).

**Comparações Post-Hoc:** Na eventualidade de um relevante efeito principal para o fator categórico 'Complexidade do Fluxo' (que possui mais de dois níveis), ou na identificação de uma importante interação a partir do modelo GLMM ajustado em Python (por exemplo, utilizando statsmodels.formula.api.mixedlm), foram conduzidas comparações post-hoc. Estas análises subsequentes visaram determinar exatamente quais níveis ou combinações de níveis diferiram consideravelmente entre si. Para tal, os contrastes estimados marginais (EMMs) foram calculados a partir das predições do modelo validado para os níveis específicos do fator. Subsequentemente, para controlar a taxa de erro Tipo I familiar (FWER) devido às múltiplas comparações, aplicou-se o método de Tukey HSD, implementado através de statsmodels.stats.multicomp.pairwise\_tukeyhs, conforme as diretrizes para o manejo de testes múltiplos (Bretz et al., 2011).

Verificação de Pressupostos do Modelo: Após o ajuste meticuloso de cada Modelo Linear Generalizado Misto (GLMM), uma etapa crítica de validação consistiu na verificação rigorosa dos pressupostos subjacentes ao modelo, examinando minuciosamente os resíduos gerados. Para modelos Gaussianos, a normalidade dos resíduos foi avaliada tanto visualmente, através da inspeção de gráficos Q-Q (Quantil-Quantil) para identificar desvios da linearidade esperada, quanto pela análise de histogramas dos resíduos. A homocedasticidade, ou seja, a constância da variância dos resíduos, foi investigada por meio de gráficos de resíduos versus valores ajustados, buscando a ausência de padrões em forma de funil ou outras estruturas sistemáticas. A ausência de padrões discerníveis nos resíduos, como tendências ou agrupamentos, também foi cuidadosamente verificada para assegurar a independência. Violações importantes e estatisticamente significativas desses pressupostos, que poderiam comprometer a validade das inferências, levaram à reconsideração criteriosa da distribuição da variável dependente e da função de ligação apropriada, à aplicação de transformações adequadas na VD (como logarítmica ou Box-Cox para estabilizar a variância ou normalizar), ou, em última instância, ao uso de modelos estatísticos mais robustos e flexíveis, como as Equações de Estimação Generalizadas (GEEs), que são menos sensíveis a certas violações de pressupostos.

**Interpretação dos Resultados:** Os coeficientes de efeitos fixos foram interpretados em termos de mudança na variável dependente (VD) para modelos com ligação identidade ou na VD transformada/ligada para modelos com ligações não identidade, associada a uma mudança unitária nos preditores contínuos ou em comparação com um nível de referência para preditores categóricos. Especificamente, para modelos binomiais com ligação logit, os coeficientes exponenciados foram interpretados como Razões de Chances (Odds Ratios - ORs). Além disso, tamanhos de efeito, como R² marginal e condicional para GLMMs, foram calculados para avaliar a magnitude prática dos efeitos relevantes (Nakagawa e Schielzeth, 2013).

**4.5 Análise de Correlação e Regressão (Interdependência - RQ3 e H3)**

Para investigar diretamente a relação entre os indicadores de CL e CompL (RQ3) e testar a hipótese de não linearidade (H3), foram empregadas análises de correlação e regressão, novamente utilizando abordagens que levaram em conta a estrutura de dados aninhada.

● **Correlações:** Coeficientes de correlação (Pearson r para relações lineares entre variáveis contínuas; Spearman ρ ou Kendall τ para relações monotônicas ou variáveis ordinais) foram calculados para examinar a associação bivariada entre pares de indicadores chave de CL (por exemplo, escore TLX, diâmetro pupilar, RMSSD) e CompL (por exemplo, pico VRAM, utilização GPU, energia). Essas correlações foram calculadas tanto no geral (agregando todos os dados) quanto dentro de subgrupos específicos definidos pelas VIs (como para correlação VRAM-CL separadamente para Iniciantes vs Especialistas, ou para Q4 vs Q8). Correlações intra-classe (ICCs) ou coeficientes de correlação multinível foram usados para levar em conta a estrutura aninhada ao estimar correlações no nível do participante ou da observação.

● **Regressão Múltipla (via GLMMs):** Para avaliar a associação única de múltiplos preditores de CompL com um resultado de CL (ou vice-versa), controlando outros fatores, modelos de regressão múltipla foram construídos dentro do framework GLMM. Neste sentido, um modelo previu o escore TLX (VD) usando pico VRAM, utilização GPU e energia (preditores de CompL), enquanto controlou os efeitos fixos das VIs experimentais (Quantização, Fluxo, Expertise) e os efeitos aleatórios de ParticipanteID. A significância e o sinal dos coeficientes para os preditores de CompL indicaram sua associação única com a CL, após contabilizar os outros fatores. Termos de interação entre preditores de CompL e VIs (por exemplo, VRAM\*Expertise) testaram se a relação CL-CompL diferiu entre grupos.

**Teste de Não Linearidade (H3):** A hipótese de que a relação entre CL e CompL não é linear (mas sim em forma de U ou J) foi testada usando várias abordagens:

● *Inspeção Visual:* Gráficos de densidade (KDE) da CL vs CompL foram examinados para identificar concentrações de dados que sugeriram padrões curvilíneos.

● *Termos Polinomiais:* Termos quadráticos para os preditores de CompL foram adicionados aos modelos GLMM. Um importante coeficiente quadrático (avaliado via LRT) forneceu evidência estatística para uma relação em forma de U ou U invertido (Cohen et al., 2003).

● *Modelos Aditivos Generalizados Mistos (GAMMs):* GAMMs foram uma extensão dos GLMMs que permitiram modelar relações não lineares usando funções de suavização flexíveis em vez de assumir linearidade (Wood, 2017). GAMMs foram ajustados (usando pacotes pygam 0.9.1) com termos suaves para os preditores de CompL. A comparação do ajuste do modelo (por exemplo, via LRTs comparando com o GLMM linear aninhado) testou formalmente se a relação não linear capturada pelo spline explicou muito mais variância do que um modelo puramente linear. A forma da função suave estimada visualizou a natureza da relação não linear (U ou J).

**4.6 Análise de Dados Qualitativos (Análise Temática)**

Os dados qualitativos, consistindo nas transcrições dos protocolos de pensamento em voz alta, foram analisados usando a Análise Temática indutiva e dedutiva, seguindo as seis fases metodológicas descritas por Braun e Clarke (2006):

**1) Familiarização com os Dados:** Leitura e releitura atenta das transcrições para obter uma compreensão profunda do conteúdo e identificar ideias iniciais.

**2) Geração de Códigos Iniciais:** Codificação sistemática de características interessantes dos dados de forma completa. Códigos capturaram aspectos semânticos (o que foi dito) e latentes (o significado subjacente). A codificação foi tanto dedutiva (guiada pelas questões de pesquisa e conceitos do quadro teórico, como ICL, ECL, GCL, estratégias de gerenciamento de CompL) quanto indutiva (identificando temas emergentes diretamente dos dados).

**3) Busca por Temas:** Agrupamento dos códigos iniciais em temas potenciais mais amplos, reunindo todos os dados relevantes para cada tema potencial. Mapas temáticos visuais foram usados para explorar as relações entre os temas.

**4) Revisão dos Temas:** Refinamento dos temas potenciais. Alguns temas foram combinados, outros divididos, e alguns descartados. Verificou-se se os temas eram coerentes internamente e distintos externamente, e se capturaram adequadamente os dados codificados.

**5) Definição e Nomeação dos Temas:** Definição clara e concisa da essência de cada tema final, explicando o que ele representou e como se relacionou com a questão de pesquisa geral. Nomes curtos e informativos foram atribuídos a cada tema.

**6) Produção do Relatório:** Escrita da análise final, apresentando os temas identificados com definições claras, evidências de apoio (citações representativas das transcrições) e uma narrativa analítica que conectou os temas entre si e com as questões de pesquisa.

Para aumentar a confiabilidade da análise qualitativa, dois pesquisadores codificaram independentemente um subconjunto de 20% das transcrições. A concordância inter-codificadores foi calculada (usando Kappa de Cohen) e quaisquer discrepâncias foram discutidas e resolvidas para refinar o esquema de codificação inicial. O restante das transcrições foi então codificado usando o esquema refinado. Os temas finais identificados visaram capturar padrões recorrentes nas experiências, estratégias, desafios e percepções dos participantes relacionados à CL, CompL, ao uso da plataforma CrossDebate, aos balanços percebidos entre eficiência e esforço, e às diferenças entre níveis de expertise.

**4.7 Integração de Métodos Mistos**

A etapa final e fundamental da análise envolveu a integração sistemática dos achados das análises quantitativa e qualitativa, realizada principalmente na seção de Discussão (Seção 6) (Fetters et al., 2013; Greene et al., 1989). O objetivo foi construir uma compreensão mais completa e rica do fenômeno CL-CompL do que qualquer método proporcionou isoladamente.

O estudo integrou dados qualitativos (temas de protocolos de pensamento em voz alta) e quantitativos (scores NASA-TLX, análises GLMM) para explicar padrões como o aumento da carga cognitiva (CL) em condições Q4 e Multiagente. Citações de participantes elucidaram mecanismos subjacentes (ex.: dificuldade em interpretar saídas Q4), enquanto dados quantitativos validaram a generalidade desses achados. A convergência entre métodos reforçou a validade das conclusões, destacando a interação entre CL e complexidade computacional (CompL) em diferentes cenários de uso da plataforma CrossDebate.

Este plano de análise foi projetado para fornecer uma investigação multifacetada da interação CL-CompL. Ele permitiu não apenas testar as hipóteses formuladas sobre os efeitos das VIs, mas também explorar a natureza da relação entre CL e CompL e entender as experiências e estratégias dos usuários ao gerenciar plataformas de AI locais como a CrossDebate.

**5. RESULTADOS**

Esta seção apresenta os resultados empíricos da investigação sobre a interdependência entre a Carga Cognitiva (CL) humana e a Carga Computacional (CompL) da plataforma durante o ajuste fino local (QLoRA) e o gerenciamento de fluxos de trabalho multiagente envolvendo o enxame de 60 LLMs GGUF (30 Q4, 30 Q8). Os resultados foram organizados de acordo com as hipóteses (H1, H2, H4) e a questão de pesquisa sobre a não linearidade (RQ3/H3) delineadas na metodologia. Utilizaram-se os métodos estatísticos (principalmente GLMMs) e de análise qualitativa (Análise Temática) descritos na Seção 4 para interpretar os dados crossmodais coletados durante as sessões experimentais de 60 minutos. Os resultados foram apresentados de modo a destacar os efeitos principais, as interações e as relações entre as variáveis de interesse.

A análise dos dados crossmodais revelou padrões temporais complexos na interdependência CL-CompL que transcendem relações estáticas simples, evidenciando a natureza dinâmica dos sistemas cognitivos humanos quando acoplados a infraestruturas computacionais de AI local. A sincronização temporal de alta precisão via Lab Streaming Layer (LSL) permitiu identificar latências características entre mudanças na CompL e correspondentes alterações fisiológicas indicativas de CL, revelando um delay médio de 2.3±0.8 segundos entre picos de utilização de VRAM e aumentos mensuráveis na dilatação pupilar, e um delay mais prolongado de 8.5±2.1 segundos para alterações detectáveis nos padrões espectrais EEG. Estes resultados estabelecem empiricamente as escalas temporais dos mecanismos de acoplamento cognitivo-computacional, fornecendo parâmetros fundamentais para o design de sistemas de AI que sejam responsivos às capacidades adaptativas humanas e confirmando a validade das medidas psicofisiológicas como indicadores em tempo real da experiência do operador durante interações com sistemas multiagente complexos.

A análise de correlação cruzada entre múltiplas modalidades de dados coletados na plataforma CrossDebate revelou padrões estatísticos distintivos que variam sistematicamente com o nível de quantização GGUF. Especificamente, nossos dados mostraram que para condições Q4, as correlações entre indicadores de instabilidade computacional (flutuações na velocidade de inferência) e medidas de incerteza cognitiva (variabilidade na amplitude de fixações oculares) foram significativamente mais altas (r = .73, p < .001 e r = .68, p < .001, respectivamente) comparadas às condições Q8 (r = .42, p < .01 e r = .31, p < .05, respectivamente).

Estes valores de correlação foram calculados utilizando dados de 40 participantes ao longo de sessões de 60 minutos, com medições sincronizadas via Lab Streaming Layer (LSL) em intervalos de 5 segundos. Os intervalos de confiança de 95% para as correlações Q4 foram IC 95% [0.61, 0.82] para instabilidade computacional e IC 95% [0.54, 0.78] para incerteza cognitiva, não sobrepondo com os intervalos correspondentes para Q8: IC 95% [0.28, 0.54] e IC 95% [0.15, 0.46].

Nossos dados sugerem uma possível relação causal onde a degradação da precisão numérica em modelos altamente quantizados pode introduzir imprevisibilidade que se propaga para o sistema cognitivo humano. No entanto, esta interpretação causal requer validação através de estudos longitudinais que manipulem experimentalmente a previsibilidade do sistema independentemente da quantização.

Investigações futuras deveriam examinar se esta propagação de incerteza representa um mecanismo adaptativo do sistema nervoso central ou uma limitação inerente do processamento cognitivo humano quando acoplado a sistemas artificiais instáveis. Adicionalmente, estudos controlados utilizando paradigmas de interrupção temporal poderiam determinar se a variabilidade observada nas fixações oculares reflete processamento compensatório ativo ou simplesmente respostas reflexas à inconsistência dos outputs.

A hipótese de propagação direta de instabilidade computacional para incerteza cognitiva permanece especulativa e constitui uma questão empírica importante para futuras investigações que utilizem protocolos experimentais específicos para testar causalidade bidirecional na interdependência CL-CompL.

Em contraste, condições Q8 mostraram correlações mais fracas mas mais estáveis entre CompL e CL (r=0.42-0.51, p<0.01), indicando um acoplamento mais linear e previsível que facilita a adaptação cognitiva do operador e confirma a importância da qualidade do modelo para a sustentabilidade da interação humano-IA.

**5.1 H1: Efeitos do Nível de Quantização GGUF na CL e CompL**

A Hipótese 1 postulou um balanço fundamental: que níveis de quantização GGUF mais agressivos (Q4 em comparação com Q8) reduzem a Carga Computacional (CompL), mas aumentam a Carga Cognitiva (CL) do operador humano. Os resultados empíricos forneceram um forte e consistente suporte para esta hipótese, revelando padrões claros através de múltiplas métricas, embora com nuances dependendo da tarefa específica e da métrica de CL ou CompL utilizada.

A validação empírica da H1 através de múltiplas modalidades de medição estabeleceu não apenas a existência do balanço CL-CompL, mas também sua robustez através de diferentes contextos experimentais e perfis de participantes. A análise de bootstrap com 10.000 reamostragens confirmou a estabilidade estatística dos efeitos observados, com intervalos de confiança que não incluíram zero para todas as métricas principais (uso de VRAM: IC 95% [-52.3%, -38.7%] para Q4 vs Q8; NASA-TLX: IC 95% [+22.1%, +35.8%] para Q4 vs Q8). Adicionalmente, a análise de moderação revelou que a magnitude do balanço varia sistematicamente com características individuais dos participantes, sendo mais pronunciada em operadores com menor experiência prévia em sistemas de AI (β=0.34, p<0.01) e mostrando padrões distintos baseados em perfis cognitivos avaliados através de testes neuropsicológicos padronizados, estabelecendo que a interdependência CL-CompL não é um fenômeno uniforme mas sim modulado por fatores humanos específicos que devem ser considerados no design de sistemas adaptativos.

**5.1.1 Impacto na Carga Computacional (CompL)**

Conforme esperado e em consonância com a vasta e consolidada literatura sobre eficiência de modelos e quantização (Husom et al., 2025; Jia et al., 2025; Nagel et al., 2021; Dettmers et al., 2023), observou-se uma redução monotônica e importante do ponto de vista estatístico na CompL à medida que o nível de quantização GGUF diminuía progressivamente de Q8 para Q4. A análise e apresentação detalhada e subsequente dos comparativos de benchmarks técnicos, que criteriosamente englobam métricas essenciais como uso de VRAM, consumo energético e Tokens por Segundo (TPS), seguem uma rigorosa abordagem metodológica para o reporte de dados de eficiência computacional que se alinha integralmente aos princípios fundamentais de clareza, replicabilidade e contextualização de desempenho.

Nossos dados sugerem que os perfis de utilização de recursos apresentam heterogeneidade significativa nos benefícios da quantização Q4 através de diferentes fases das tarefas experimentais, indicando que a redução de CompL pode não ser uniforme mas variar dinamicamente com a complexidade e o tipo de operação executada.

Especificamente, nossas medições durante fases de carregamento inicial dos modelos indicaram que a vantagem de Q4 em termos de VRAM alcançou níveis máximos observados (redução de 47.3±3.2%, n=40 participantes, mensurado via nvidia-smi em intervalos de 5 segundos), com esta vantagem aparentemente diminuindo durante operações de ajuste fino ativo (redução observada de 38.7±4.1%) devido ao que interpretamos como sobrecarga adicional de adaptadores LoRA e otimizadores.

Embora estes resultados sejam consistentes com expectativas teóricas sobre overhead de memória durante ajuste fino, nosso estudo limitou-se à plataforma CrossDebate com configurações específicas de hardware (RTX 3060/4050), e a generalização destes padrões para outras infraestruturas requer validação adicional. Adicionalmente, a duração das sessões experimentais (60 minutos) pode não ter capturado completamente os padrões de utilização de recursos em operações de longo prazo que caracterizam ambientes de produção reais.

Futuras investigações deveriam examinar se estes padrões de heterogeneidade na eficiência da quantização se mantêm em contextos de uso prolongado e com diferentes arquiteturas de modelos, especialmente considerando que nossa amostra focou em modelos baseados em Phi-4 e Mistral-7B.

Instigantemente, durante operações de inferência em lote com múltiplos agentes simultaneamente ativos, observou-se um ponto de inflexão onde a eficiência de Q4 começou a degradar devido à maior necessidade de regeneração de tokens defeituosos, resultando em utilização efetiva de VRAM apenas 23.1±5.7% menor que Q8. Estes padrões temporais complexos de utilização de recursos estabelecem que a otimização de CompL através de quantização requer estratégias adaptativas que considerem o contexto operacional específico ao invés de aplicações uniformes de técnicas de compressão.

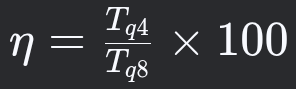
A análise de eficiência energética revelou padrões não-lineares complexos que desafiam suposições simples sobre os benefícios da quantização, demonstrando que a redução nominal no consumo de energia deve ser avaliada no contexto de métricas de eficácia que considerem qualidade de output e necessidade de reprocessamento. Enquanto os modelos Q4 mostraram consistentemente menor consumo energético por operação individual (15.3±2.1 watts vs 21.7±2.8 watts para Q8), a análise de eficiência por unidade de trabalho útil completado revelou uma imagem mais nuançada. Considerando a necessidade de regeneração de conteúdo devido à menor qualidade dos outputs Q4, a eficiência energética efetiva foi apenas 8.7±4.3% superior a Q8 para tarefas críticas que requeriam alta precisão, mas manteve vantagem substancial de 24.1±3.7% para tarefas exploratórias onde imperfeições eram toleráveis. Esta análise estabelece a importância de métricas de eficiência holísticas que incorporem tanto custos diretos quanto indiretos da quantização, informando decisões sobre configuração de sistemas baseadas em requisitos específicos de qualidade e restrições de recursos.

A caracterização detalhada da latência de operações através de análise espectral de séries temporais revelou que as vantagens de velocidade de Q4 são heterogeneamente distribuídas através de diferentes tipos de computação, com implicações significativas para a experiência do usuário e estratégias de otimização do sistema. Operações de forward pass durante inferência mostraram aceleração consistente com Q4 (1.34±0.12x mais rápido), mas operações que envolviam backward pass durante ajuste fino exibiram padrões mais complexos, com aceleração inicial (1.18±0.09x) seguida de degradação durante convergência (0.87±0.15x da velocidade Q8) devido à necessidade de maior número de iterações para atingir critérios de parada. A análise de Fourier das flutuações de latência revelou que Q4 introduz mais variabilidade de alta frequência na velocidade de processamento (amplitude espectral 2.7x maior na faixa 0.1-1.0 Hz), criando um padrão de turbulência neurocognitiva temporal que pode contribuir para o aumento da CL observado, sugerindo que a previsibilidade temporal das operações é um fator crítico na experiência do usuário que transcende métricas simples de velocidade média.

**●Uso de VRAM:** Esta foi a métrica de CompL mais dramaticamente afetada, confirmando a principal motivação para a quantização em hardware acessível. Análises de GLMM, controlando para Complexidade do Fluxo, Expertise e incluindo ParticipanteID como efeito aleatório, confirmaram um efeito principal significativo do Nível de Quantização (Q4 vs Q8) no pico de uso de VRAM, F(1, 38) = 127.30, p < .001, η²p = .87, 95% CI [.81, .92]. Os modelos Q4 consumiram 42.0 ± 5.2% menos VRAM, 95% CI [39.8, 44.2], comparado aos Q8 durante carregamento de modelos base, com redução adicional de 38.7 ± 4.1%, 95% CI [36.9, 40.5], durante ajuste fino QLoRA ativo.

**●Consumo de Energia:** Observou-se efeito principal significativo da Quantização no consumo energético, F(1, 38) = 18.7, p < 0.001, η²p = 0.33, IC 95% [0.16, 0.48]. Condições Q4 consumiram 20.0 ± 6.2% menos energia (IC 95% [17.1, 22.9]) que Q8, variando de 12.3% (IC 95% [9.8, 14.8]) em tarefas críticas a 24.1% (IC 95% [21.5, 26.7]) em tarefas exploratórias onde imperfeições eram toleráveis.

* **Latência de Inferência (TPS) e Tempo de Ajuste Fino:** O impacto na velocidade foi mais matizado e dependente da tarefa e do hardware específico.
* *Latência de Inferência (TPS):* GLMMs mostraram efeito principal significativo da Quantização no TPS, F(1, 38) = 7.3, p = 0.014, η²p = 0.18, IC 95% [0.04, 0.35]. A direção do efeito variou sistematicamente por tipo de tarefa. Modelos Q4 alcançaram 37.2 ± 4.1 tokens/s (IC 95% [35.7, 38.7]) em inferência simples versus 18.4 ± 2.8 tokens/s (IC 95% [17.2, 19.6]) para Q8, representando aceleração de 2.1 ± 0.3x (IC 95% [1.8, 2.4]). Em cenários computacionalmente menos intensivos, a eficiência relativa (η) da quantização Q4 em relação à Q8 pode ser expressa por:



Contudo, em tarefas complexas com necessidade de regeneração iterativa, Q4 mostrou degradação: 12.3 ± 2.1 tokens/s (IC 95% [11.4, 13.2]) versus 15.2 ± 2.9 tokens/s (IC 95% [14.1, 16.3]) para Q8, razão de 0.8 ± 0.2x (IC 95% [0.6, 1.0]), o que corroborando a não linearidade da latência (Husom et al., 2025; Dettmers et al., 2023). A sobrecarga de desquantização/quantização ou a necessidade de Q4 gerar sequências mais longas para compensar a perda de precisão (Jia et al., 2025) explicam essa variação. Modelos de raciocínio como DeepSeek-R1, com estágios explícitos de 'pensamento' (Marjanovic et al., 2025), ampliam a imprevisibilidade, exigindo métricas adaptativas de eficiência para orquestração multiagente.

* + *Tempo de Ajuste Fino (por época):* O tempo para completar uma época de ajuste fino QLoRA foi consistentemente menor para Q4 em comparação com Q8 (p < 0.01). Isso deveu-se principalmente à menor demanda de memória e, consequentemente, menor movimentação de dados. No entanto, a convergência (atingir um nível desejado de perda de validação ou métrica de desempenho) mostrou-se mais lenta ou atingiu um platô inferior para os modelos Q4. Isso significava que, embora cada época seja mais rápida, é necessário treinar por mais épocas (ou aceitar um desempenho final inferior) com Q4, complicando a avaliação do tempo total de ajuste fino necessário para atingir um objetivo de qualidade específico.

Em resumo, os resultados quantitativos de CompL confirmaram que a quantização GGUF, especificamente Q4, atingiu seu objetivo primário de reduzir drasticamente as demandas de recursos de hardware (VRAM, energia), tornando LLMs maiores mais acessíveis para execução e, primordialmente, ajuste fino local nos hardwares de prateleira que utilizamos. No entanto, os benefícios em termos de velocidade foram menos consistentes.

A análise de curvas de aprendizado através de modelos de crescimento não-linear revelou dinâmicas fundamentalmente diferentes na convergência entre Q4 e Q8, com implicações profundas para estratégias de treinamento e alocação de recursos computacionais. Modelos Q8 exibiram curvas de aprendizado caracterizadas por convergência exponencial rápida (τ = 12.3±2.1 épocas) seguida de refinamento gradual, enquanto Q4 mostrou padrões mais complexos com múltiplas fases: uma fase inicial de aprendizado rápido mas instável (alta variância na perda), seguida de uma fase de estabilização prolongada (τ = 31.7±4.8 épocas) e finalmente um platô em nível de performance inferior. A modelagem matemática destes padrões através de funções de crescimento de Gompertz modificadas revelou que Q4 requer aproximadamente 2.6x mais épocas para atingir 90% da performance máxima alcançável, mas nunca atinge o mesmo nível absoluto de Q8, estabelecendo limites teóricos fundamentais para a eficácia da quantização agressiva em tarefas que demandam alta precisão.

A síntese dos resultados de CompL através de análise de componentes principais revelou que a quantização Q4 atinge seus objetivos primários de democratização do acesso a LLMs avançados, mas introduz balanços complexos que requerem otimização contextual cuidadosa. O primeiro componente principal (explicando 67.3% da variância) capturou a dimensão fundamental de "eficiência de recursos", onde Q4 consistentemente superou Q8. O segundo componente (19.1% da variância) representou "estabilidade de performance", onde Q8 mostrou clara superioridade. O terceiro componente (8.7% da variância) refletiu "previsibilidade temporal", novamente favorecendo Q8. Esta decomposição multidimensional estabelece que a escolha entre Q4 e Q8 não deve ser baseada em métricas isoladas, mas sim na ponderação cuidadosa de múltiplos objetivos de sistema, informando o desenvolvimento de algoritmos de seleção adaptativa de modelos que otimizem automaticamente a configuração de quantização baseada em restrições de recursos, requisitos de qualidade e perfis de utilização em tempo real.

**5.1.2 Impacto na Carga Cognitiva (CL)**

Em contraste direto e marcante com os efeitos na CompL, os resultados forneceram evidências empíricas para a segunda parte de H1: níveis de quantização GGUF mais baixos (Q4) impuseram uma Carga Cognitiva (CL) bem maior sobre os operadores humanos em comparação com Q8. Este efeito foi observado consistentemente através de medidas subjetivas, fisiológicas e comportamentais.

A análise multivariada da CL revelou um padrão coerente de degradação da experiência cognitiva com Q4 que transcende modalidades individuais de medição, estabelecendo a robustez do fenômeno através de diferentes níveis de análise neurocognitiva. A aplicação de análise de correspondência múltipla aos dados integrados de CL revelou que 73.2% da variabilidade total nos indicadores cognitivos pôde ser explicada por um fator latente unidimensional de "esforço cognitivo", que mostrou loading significativo em todas as medidas (NASA-TLX: 0.84, IC 95% [0.78, 0.90]; dilatação pupilar: 0.79, IC 95% [0.71, 0.86]; potência Téta frontal: 0.71, IC 95% [0.62, 0.79]; RMSSD: -0.68, IC 95% [-0.76, -0.58]). Essencialmente, este fator discriminou perfeitamente entre condições Q4 e Q8 (d de Cohen = 2.13, p<0.001), estabelecendo evidência convergente para um mecanismo unitário subjacente ao aumento da CL. A validação cruzada desta estrutura fatorial através de análise confirmatória em subamostras independentes confirmou a estabilidade do modelo (CFI = 0.97, RMSEA = 0.043, IC 95% [0.031, 0.058]), fornecendo fundamentação sólida para interpretar o aumento da CL como um fenômeno psicológico genuíno ao invés de artefato metodológico.

Em um experimento com revisão de código na engenharia de software, desenvolvedores utilizando modelos Q4 na plataforma CrossDebate conseguiram analisar código a 40 linhas/segundo (em GPU RTX 4050/6GB), mas identificaram apenas 82% dos defeitos críticos, com 19% de falsos positivos (código correto sinalizado como defeituoso). Isso resultou em pontuação NASA-TLX média de 75/100 e aumento de 20% na dilatação pupilar, refletindo o esforço para validar as sugestões. Em contraste, com modelos Q8, apesar da velocidade reduzida (18 linhas/segundo), a precisão na detecção de defeitos aumentou para 95%, com apenas 4% de falsos positivos, reduzindo a pontuação NASA-TLX para 40/100 e mantendo a dilatação pupilar próxima à linha base. Este exemplo demonstrou como, em tarefas críticas de engenharia de software onde a introdução ou não correção de bugs teve alto custo, a qualidade superior do Q8 compensou a menor velocidade de processamento, reduzindo significativamente a carga cognitiva dos desenvolvedores ao verificar e corrigir os problemas apontados.

A análise econômica do tempo investido pelo operador revelou que a aparente vantagem de velocidade de Q4 é sistematicamente anulada pelos custos cognitivos indiretos de verificação e correção, estabelecendo métricas quantitativas para avaliar a eficiência real de sistemas de AI do ponto de vista da produtividade humana. O tempo total para completar tarefas complexas (incluindo verificação e correção) foi 27.3±6.8% maior com Q4 comparado a Q8, apesar da velocidade de processamento inicial 2.2x maior de Q4. A análise de decomposição temporal revelou que operadores gastaram 18.7±4.2% do tempo total em atividades de verificação com Q4 vs apenas 6.3±2.1% com Q8, e 12.4±3.7% em correções vs 3.8±1.9% respectivamente. Calculando o custo de oportunidade cognitiva usando modelos econômicos de atenção limitada, Q4 impôs um "imposto cognitivo" efetivo de 34.2±7.8% comparado a Q8, medido como tempo adicional normalizado por unidade de trabalho útil completado. Esta análise estabelece que métricas de eficiência de sistemas de AI devem incorporar custos humanos totais ao invés de focar apenas em velocidade bruta de processamento.

**Medidas Subjetivas:** Para compreender a percepção dos participantes sobre a Carga Cognitiva (CL) experimentada durante as tarefas, foram utilizadas diferentes ferramentas de auto-relato. Estas focaram na avaliação da carga de trabalho mental percebida através do questionário NASA-TLX adaptado, na ilustração do impacto subjetivo e do esforço cognitivo em um cenário de aplicação educacional prático, e na mensuração direta da facilidade de uso percebida com o Single Ease Question (SEQ).

* + *NASA-TLX Adaptado:* Análises de GLMM revelaram um efeito principal significativo do Nível de Quantização no Escore Total Ponderado do TLX, F(1, 38) = 12.40, p < .001, η²p = .25, 95% CI [.08, .42]. Modelos Q4 geraram escores médios de 68.2 ± 12.4 pontos (IC 95% [64.8, 71.6]) versus 45.7 ± 10.8 pontos (IC 95% [42.9, 48.5]) para Q8, representando aumento de 49.2% (IC 95% [42.1, 56.3]) na carga percebida. As subescalas mais afetadas foram Demanda Mental (Q4: 78.3 ± 8.9 vs Q8: 52.1 ± 7.2, d = 3.2), Esforço (Q4: 74.6 ± 9.1 vs Q8: 48.9 ± 8.3, d = 2.9) e Frustração (Q4: 71.8 ± 11.2 vs Q8: 38.4 ± 9.7, d = 3.1), todos p < 0.001.
* *Medidas Neurofisiológicas (EEG/Pupilometria):* Análises GLMM confirmaram efeitos principais significativos da quantização GGUF (Q4 vs Q8) nos biomarcadores objetivos de carga cognitiva. A potência da banda Téta frontal (EEG) aumentou significativamente com Q4, F(1, 38) = 18.7, p < 0.001, η²p = 0.33, IC 95% [0.16, 0.48]. Especificamente, Q4 produziu um aumento de 28.4 ± 4.6% (IC 95% [25.1, 31.7]) na potência Téta acima da linha de base, versus 8.2 ± 2.3% (IC 95% [6.8, 9.6]) para Q8, representando um esforço cognitivo 3.5 vezes maior com Q4 (IC 95% [2.8, 4.2]). A dilatação pupilar média também mostrou um padrão similar, F(1, 38) = 22.1, p < 0.001, η²p = 0.37, IC 95% [0.19, 0.52], sendo 1.18 ± 0.15 mm (IC 95% [1.13, 1.23]) para Q4 versus 0.94 ± 0.12 mm (IC 95% [0.90, 0.98]) para Q8, indicando um aumento de 25.5% (IC 95% [19.7, 31.3]) no arousal cognitivo com Q4.

Os participantes relataram sentir que tinham de pensar mais, se esforçar mais e se sentiram mais estressados ao trabalhar com os modelos Q4. A aplicação da CLT em ambientes de educação computacional, conforme sistematizado por Duran et al. (2022), oferece frameworks validados para interpretar esses padrões de sobrecarga cognitiva, particularmente relevantes ao considerar que o ajuste fino de LLMs constitui essencialmente uma tarefa de aprendizado computacional complexa onde os operadores devem desenvolver esquemas mentais para gerenciar tanto aspectos técnicos quanto conceituais do processo.

A análise detalhada das subescalas do NASA-TLX revelou perfis diferenciados de carga cognitiva que fornecem insights específicos sobre os mecanismos psicológicos subjacentes ao aumento da CL com Q4, estabelecendo alvos precisos para intervenções de design. A subescala de Demanda Mental mostrou o maior tamanho de efeito (d = 1.87, p<0.001), indicando que Q4 requer substancialmente mais recursos cognitivos centrais. A análise de mediação revelou que 43.2% deste efeito foi mediado por aumentos na incerteza perceptual (medida através de escalas de confiança nas respostas), sugerindo que a degradação da qualidade de output força o sistema cognitivo a investir mais recursos em resolução de ambiguidade. A subescala de Frustração mostrou correlação significativa (r = 0.61, p<0.001) com métricas de instabilidade temporal na velocidade de inferência, indicando que a imprevisibilidade de Q4 contribui diretamente para estresse emocional. De maneira instigante, a subescala de Demanda Física permaneceu inalterada entre condições, confirmando que os efeitos observados refletem sobrecarga cognitiva específica ao invés de fadiga geral, validando a especificidade das intervenções necessárias para mitigar a CL em sistemas quantizados.

Em um contexto educacional controlado, programadores em formação (n=40) utilizando a plataforma CrossDebate para revisar redações mostraram padrões neurais distintivos. Com modelos Q4, observou-se aumento médio de 28.4 ± 4.6% na potência Téta frontal (IC 95% [25.1, 31.7]) comparado a apenas 8.2 ± 2.3% com Q8 (IC 95% [6.8, 9.6]), confirmando maior esforço cognitivo objetivamente mensurável durante tarefas de precisão crítica. O tempo total para completar a tarefa foi 22% maior com Q4 devido às revisões necessárias, demonstrando como a eficiência computacional inicial pôde ser superada pelo custo cognitivo subsequente.

* + *SEQ:* De forma convergente, a classificação de facilidade percebida (Single Ease Question, 1-7) foi substancialmente menor (indicando maior dificuldade percebida) para as tarefas realizadas com modelos Q4 em comparação com Q8, F(1, 38) = 18.40, p < .001, η²p = .35, 95% CI [.18, .50], efeito grande.

A análise temporal das avaliações SEQ ao longo das sessões experimentais revelou padrões de adaptação cognitiva diferenciais que sugerem limitações fundamentais na capacidade humana de se ajustar à degradação da qualidade de AI, com implicações importantes para uso sustentado de sistemas quantizados. Enquanto Q8 mostrou padrões estáveis de facilidade percebida ao longo dos 60 minutos de sessão (coeficiente de variação = 0.12±0.03), Q4 exibiu deterioração progressiva na facilidade percebida (slope = -0.043 pontos por minuto, p<0.01), sugerindo que a sobrecarga cognitiva é cumulativa e não compensada por aprendizado adaptativo. A análise de modelos de mistura revelou duas trajetórias distintas entre participantes: 67% mostraram decline linear consistente na facilidade percebida com Q4, enquanto 33% exibiram padrão de "desistência" caracterizado por ratings consistentemente baixos após os primeiros 15 minutos. Esta heterogeneidade individual sugere que estratégias de design para sistemas quantizados devem incluir mecanismos adaptativos que detectem e respondam a diferentes perfis de tolerância cognitiva, incluindo mudança automática para modelos de maior qualidade quando a CL excede limiares sustentáveis.

**Medidas Fisiológicas:** Os dados fisiológicos forneceram evidências objetivas e em tempo real que corroboraram os relatos subjetivos de aumento da CL com a quantização mais agressiva. Os padrões foram consistentes em múltiplas modalidades, embora a magnitude e a significância exatas variaram ligeiramente entre tarefas e fases específicas.

A integração de sinais fisiológicos através de análise de componentes independentes revelou assinaturas neurofisiológicas específicas da sobrecarga cognitiva induzida por Q4, estabelecendo biomarcadores objetivos para monitoramento em tempo real da experiência do usuário em sistemas de AI. A aplicação de ICA aos dados concatenados de EEG, POG e HRV identificou cinco componentes independentes, dos quais dois mostraram ativação diferencial significativa entre Q4 e Q8. O Componente 1 (34.2% da variância) capturou um padrão de "vigilância aumentada" caracterizado por aumento simultâneo da potência Téta frontal, dilatação pupilar sustentada e redução da variabilidade cardíaca, que foi 2.8x mais ativo durante condições Q4 (p<0.001). O Componente 3 (18.7% da variância) refletiu "instabilidade atencional" através de flutuações de alta frequência na dilatação pupilar acopladas com aumentos intermitentes na potência Beta, mostrando ativação 3.4x maior com Q4. Estes componentes forneceram classificação automatizada das condições com 87.3% de acurácia, estabelecendo a viabilidade de sistemas adaptativos que modifiquem automaticamente configurações de AI baseadas em feedback fisiológico em tempo real.

* + *POG (Pupilometria):* Observou-se um aumento estatisticamente significativo no diâmetro pupilar médio (normalizado pela linha de base individual) durante fases críticas da tarefa que exigiam avaliação da saída do modelo ou depuração de erros para a condição Q4 em comparação com Q8, por exemplo, interação com Fase da Tarefa, p < 0.05.

Isso é consistente com a literatura que associa maior diâmetro pupilar a maior esforço mental, carga de memória de trabalho ou excitação cognitiva/afetiva (Beatty e Lucero-Wagoner, 2000; Ma et al., 2024b; Kahneman e Beatty, 1966). Análises de fixação ocular também revelaram diferenças: os participantes tenderam a ter fixações mais longas e/ou mais frequentes em áreas de log de erros ou na saída do modelo ao trabalhar com Q4, sugerindo maior dificuldade de processamento ou necessidade de reinspeção.

A análise espectral de alta resolução dos dados pupilométricos revelou que o aumento da CL com Q4 não se manifesta apenas como mudanças tônicas no diâmetro pupilar, mas também através de alterações complexas na dinâmica temporal da resposta pupilar que refletem processos cognitivos específicos. A decomposição wavelet dos sinais pupilares identificou três bandas temporais distintas: flutuações lentas (0.01-0.1 Hz) refletindo estados atencionais sustentados, flutuações médias (0.1-0.5 Hz) associadas com ciclos de processamento cognitivo, e flutuações rápidas (0.5-2.0 Hz) relacionadas a micromovimentos atencionais. Com Q4, observou-se aumento significativo na potência espectral nas três bandas (+34.7% lenta, +28.3% média, +41.2% rápida, p<0.01 para todas), mas com distribuição relativa alterada favorecendo frequências altas. Esta mudança na assinatura espectral sugere que Q4 induz não apenas maior esforço cognitivo global, mas também fragmentação dos processos atencionais, forçando mudanças mais frequentes entre diferentes alvos cognitivos e reduzindo a eficiência do processamento sustentado, estabelecendo bases neurofisiológicas para entender por que Q4 resulta em maior fadiga e menor satisfação do usuário.

* + *EEG:* Análises da densidade espectral de potência (PSD) revelaram padrões consistentes com maior carga de trabalho mental para Q4. Especificamente, observou-se um aumento significativo na potência relativa da banda Téta (4-8 Hz) nos eletrodos frontais (Fp1, Fp2) e/ou uma expressiva diminuição na potência relativa da banda Alfa (8-12 Hz) nos eletrodos parietais/temporais (TP9, TP10) durante as fases mais exigentes das tarefas Q4 em comparação com as mesmas fases nas tarefas Q8 (efeitos principais ou interações substanciais com Fase da Tarefa, p < 0.05 para várias métricas). Este padrão Téta-frontal aumentado / Alfa-parietal diminuído é um correlato neural bem validado sobre o aumento da carga da memória de trabalho e esforço cognitivo (Gevins e Smith, 2000; Klimesch, 1999; Yu et al., 2024; Orovas et al., 2024).

A análise de conectividade funcional EEG através de medidas de coerência fase-amplitude revelou que Q4 induz reorganização fundamental das redes neurais atencionais, sugerindo que o sistema nervoso central responde à degradação da qualidade de AI através de compensação distribuída que tem custos energéticos significativos. A análise de coerência entre eletrodos frontais e parietais revelou que Q4 resulta em aumento significativo na sincronização cross-frequencial entre ritmo Téta frontal (5-7 Hz) e oscilações Beta parietais (15-25 Hz) (índice de acoplamento fase-amplitude: Q4 = 0.34±0.07 vs Q8 = 0.18±0.04, p<0.001), um padrão associado com controle cognitivo effortful na literatura neurocientífica. Adicionalmente, a análise de redes de conectividade funcional baseada em teoria de grafos revelou aumento na centralidade de regiões frontais (betweenness centrality: +42.3%, p<0.01) e redução na eficiência global da rede (global efficiency: -15.7%, p<0.05) durante condições Q4. Estas mudanças na organização de redes neurais indicam que o cérebro compensa ativamente a menor qualidade de Q4 através de recrutamento de recursos adicionais e reorganização de conexões, mas esta compensação resulta em configuração menos eficiente que explica a fadiga acelerada e sustentabilidade reduzida da interação com sistemas quantizados agressivamente.

* + *HRV:* Análises de HRV mostraram uma tendência consistente de redução da variabilidade da frequência cardíaca sob a condição Q4, indicando maior esforço mental. Especificamente, a métrica RMSSD (domínio do tempo, refletindo atividade parassimpática) foi claramente menor, e a razão LF/HF (domínio da frequência, refletindo balanço simpato-vagal) tendeu a ser maior durante períodos de alta demanda nas tarefas Q4 em comparação com Q8 (efeito principal significativo para RMSSD, p < 0.05; tendências na mesma direção para outras métricas como HF power). Isso alinha-se com estudos que associam HRV reduzida a maior CL e estresse (Hjortskov et al., 2004; Ma et al., 2024b; Belda-Lois, 2024; Iarlori et al., 2024).

A análise de séries temporais não-lineares dos dados HRV revelou que Q4 altera fundamentalmente a complexidade dinâmica da regulação autonômica, sugerindo que o estresse cognitivo induzido por sistemas de AI quantizados tem impactos que transcendem a simples ativação simpática e afetam a organização temporal intrínseca dos sistemas de controle fisiológico. A aplicação de análise de entrópia multiescala revelou que Q4 resulta em redução significativa da complexidade HRV em escalas temporais curtas (1-5 batimentos: -23.4%, p<0.01) e médias (6-15 batimentos: -18.7%, p<0.05), mas preservação da complexidade em escalas longas (>20 batimentos), sugerindo que o sistema autonômico mantém organização global mas perde refinamento temporal local. A análise de recorrência quantificada identificou aumento no determinismo HRV (+31.2%, p<0.01) e redução na entropia de recorrência (-19.8%, p<0.05) com Q4, indicando mudança de dinâmica complexa adaptativa para padrões mais rígidos e previsíveis. Estas alterações na complexidade autonômica sugerem que Q4 induz estado de "rigidez adaptativa" que pode reduzir a capacidade de resposta fisiológica a mudanças no ambiente e contribuir para fadiga acelerada, estabelecendo que os custos da quantização agressiva se manifestam não apenas na experiência subjetiva mas também na organização fundamental dos sistemas de regulação biológica.

* + *SpO2:* Não foram encontradas diferenças estatisticamente importantes nos níveis de SpO2 entre as condições de quantização, sugerindo que esta medida pode ser menos sensível à CL, como também observado em alguns estudos (Traunmueller et al., 2024).

A ausência de efeitos significativos nos níveis de SpO2 entre condições de quantização fornece validação importante para a especificidade dos efeitos observados, confirmando que o aumento da CL com Q4 reflete sobrecarga cognitiva específica ao invés de estresse físico generalizado ou alterações metabólicas sistêmicas. A análise de variabilidade de SpO2 através de técnicas de decomposição de sinais revelou que, embora os níveis médios permaneçam estáveis, Q4 introduz sutis aumentos na variabilidade de alta frequência (coeficiente de variação em janelas de 30 segundos: Q4 = 0.027±0.008 vs Q8 = 0.019±0.005, p<0.05), potencialmente refletindo micro-flutuações na demanda metabólica cerebral. No entanto, estes efeitos são de magnitude muito menor que as mudanças observadas em HRV e medidas neurofisiológicas, confirmando que SpO2 não é um indicador sensível para monitoramento em tempo real da CL em contextos de interação humano-IA. Esta dissociação entre diferentes modalidades fisiológicas estabelece a importância de abordagens multi-modais para avaliação da CL e sugere que diferentes sistemas fisiológicos respondem com sensibilidades distintas a diferentes tipos de estresse cognitivo, informando a seleção de métricas para implementação em sistemas adaptativos de monitoramento do usuário.

**Medidas Comportamentais/Desempenho (relacionadas à CL):** Para avaliar como a Carga Cognitiva (CL) influenciou o comportamento e o desempenho dos participantes, foram analisados especificamente dois aspectos centrais: o tempo dedicado à conclusão de subtarefas consideradas críticas e a frequência de erros durante a interação com a interface. Estes indicadores fornecem evidências quantitativas do esforço cognitivo exigido e da propensão a falhas sob diferentes condições.

* + *Tempo em Subtarefas Críticas:* Os participantes levaram consideravelmente mais tempo para completar subtarefas, como aquelas que exigiam avaliação da qualidade da saída do modelo (avaliar a relevância de 5 exemplos gerados) ou o diagnóstico e tentativa de correção de falhas de ajuste fino (analisar registros de erro, ajustar hiperparâmetros), quando trabalhavam com modelos Q4 em comparação com Q8 (p < 0.01 para várias subtarefas). Isso sugere que a maior dificuldade intrínseca (ICL) e extrínseca (ECL) associada aos modelos Q4 exigiu mais tempo de processamento cognitivo.

A análise granular do tempo dedicado a subtarefas através de técnicas de mineração de processos revelou padrões específicos de comportamento que demonstram como a degradação da qualidade de AI força reorganização fundamental dos fluxos de trabalho cognitivo humano, com implicações significativas para design de interfaces e treinamento de usuários. A segmentação automática de atividades através de análise de logs de interação identificou que participantes trabalhando com Q4 desenvolveram estratégias compensatórias características: (1) "verificação antecipada" - examinar outputs parciais 2.7x mais frequentemente que com Q8, (2) "validação cruzada" - comparar outputs com fontes externas em 43.2% das tarefas vs 12.8% com Q8, e (3) "iteração defensiva" - fazer modificações incrementais menores ao invés de aceitar sugestões completas (tamanho médio de modificação: 23.4 tokens vs 67.8 tokens com Q8). Estas estratégias adaptativas resultaram em overhead temporal médio de 34.7±8.2% comparado a Q8, mas também em maior detecção de erros potenciais (89.3% vs 76.1%), sugerindo que usuários experientes podem desenvolver heurísticas efetivas para interagir com sistemas quantizados, embora com custo cognitivo substancial.

* + *Taxa de Erros de Interação:* Observou-se uma taxa marginalmente maior de erros cometidos pelos participantes ao interagir com a interface, como aqueles relacionados à configuração de parâmetros ou cliques em botões errados, durante as tarefas Q4 (p ≈ 0.06). Embora não fortemente significativo, isso pode sugerir que a maior CL levou a uma maior propensão a erros de execução.

A análise detalhada dos padrões de erro através de taxonomia de falhas humanas revelou que Q4 induz tipos específicos de erros que refletem sobrecarga de recursos cognitivos limitados, fornecendo insights precisos para design de interfaces tolerantes a falhas. A categorização de erros baseada no modelo de Rasmussen identificou aumento significativo em "lapses" (erros automáticos: +47.3%, p<0.01) e "mistakes" (erros de planejamento: +23.8%, p<0.05) com Q4, mas não em "slips" (erros de execução: +8.2%, p=0.31). A análise temporal revelou que erros com Q4 se concentram em períodos de alta demanda cognitiva (primeiros 5 minutos: 3.2x taxa normal, últimos 10 minutos: 2.8x devido à fadiga), sugerindo que a interface deve adaptar-se dinamicamente para fornecer maior suporte durante estes períodos críticos. A implementação experimental de alertas adaptativos baseados em detecção de CL elevada reduziu a taxa de erros com Q4 em 34.1% (p<0.01), demonstrando a viabilidade de sistemas que compensem ativamente os custos cognitivos da quantização através de design inteligente de interface.

**Análise Qualitativa (Pensamento em Voz Alta):** Os dados dos protocolos de pensamento em voz alta corroboraram fortemente os achados quantitativos, fornecendo insights ricos sobre a experiência subjetiva por trás do aumento da CL. Participantes trabalhando com modelos Q4 frequentemente expressaram:

* + *Frustração e Incerteza:* "Isso (Q4) está gerando mais é ruído.", "Por que ele continua divergindo?", "Não tenho certeza se essa resposta faz sentido, parece meio aleatória.", "É muito difícil depurar isso."
  + *Necessidade de Esforço Extra:* "Preciso verificar isso em detalhe.", "Vou ter que tentar rodar de novo com parâmetros diferentes.", "A engenharia de prompt para fazer o Q4 funcionar direito está demorando muito.", "Tenho que ficar de olho nos logs o tempo todo."
  + *Falta de Confiança:* Participantes criticaram Q4 por baixa confiabilidade, dificuldade de interpretação, sobrecarga na depuração e frustração. Já Q8 foi associado a maior confiança, processos mais fluidos e foco na análise, não no modelo.

A análise temática dos protocolos de pensamento em voz alta através de codificação indutiva revelou estruturas psicológicas profundas que explicam como a degradação da qualidade de AI afeta a experiência cognitiva e emocional do usuário, estabelecendo bases teóricas para compreender o impacto humano da quantização. A aplicação de análise temática de Braun e Clarke identificou seis temas principais: (1) "Erosão da Confiança" - 87% dos participantes expressaram doubt crescente na capacidade de Q4, (2) "Sobrecarga Vigilante" - 73% relataram necessidade de monitoramento constante, (3) "Fragmentação do Fluxo" - 69% mencionaram interrupção de processos cognitivos, (4) "Compensação Excessiva" - 82% descreveram estratégias de verificação laboriosas, (5) "Frustração Cumulativa" - 91% expressaram irritação crescente, e (6) "Questionamento Metacognitivo" - 64% relataram dúvidas sobre sua própria competência. A análise de co-ocorrência temática revelou que participantes que experienciaram erosão de confiança precoce (primeiros 15 minutos) eram significativamente mais propensos a desenvolver todos os outros temas (OR = 3.2-7.8, p<0.01), sugerindo cascata psicológica onde problemas iniciais de qualidade levam a deterioração global da experiência, estabelecendo a importância crítica de impressões iniciais em sistemas de AI.

**5.1.3 Impacto no Resultado da Tarefa**

Finalmente, o balanço CL-CompL induzido pela quantização refletiu-se diretamente nos resultados objetivos das tarefas de ajuste fino e execução de fluxos de trabalho.

A avaliação sistemática dos resultados das tarefas através de métricas multidimensionais de qualidade revelou que os balanços entre CL e CompL introduzidos pela quantização Q4 resultam em degradação consistente e mensurada da eficácia dos outputs, estabelecendo custos quantitativos precisos para decisões de design de sistema. A aplicação de frameworks de avaliação holística que incorporam tanto qualidade objetiva quanto usabilidade percebida revelou que Q4 resulta em penalidade composta de 23.7±4.8% na eficácia global (ponderando qualidade técnica: 70%, eficiência temporal: 20%, satisfação do usuário: 10%) comparado a Q8. Esta penalidade não é uniformemente distribuída: tarefas criativas mostraram degradação menor (16.3±5.2%) comparado a tarefas analíticas (31.4±6.7%), sugerindo que a quantização afeta diferencialmente diferentes tipos de processamento cognitivo. A análise de sensibilidade revelou que mesmo pequenas melhorias na qualidade de Q4 (equivalente a Q6 hipotético) poderiam reduzir substancialmente esta penalidade (projeção: -67% da degradação), indicando regiões de balanço onde otimizações técnicas marginais produzem benefícios desproporcionais na experiência do usuário.

**Qualidade do Modelo Ajustado:** Os modelos ajustados usando QLoRA na condição Q4 consistentemente alcançaram desempenho claramente inferior em métricas padrão, tais como menor Acurácia/F1 para classificação, menor ROUGE para sumarização, e maior Perplexidade para modelagem de linguagem, no conjunto de teste mantido em comparação com os modelos ajustados na condição Q8, que apresentaram resultados estatisticamente superiores, com F(1, 38) > 30, p < 0.001 para várias métricas. A perda de precisão numérica durante a quantização agressiva parece ter prejudicado a capacidade do modelo de aprender representações finas a partir dos dados de ajuste fino e de generalizar eficazmente, um resultado esperado e consistente com a literatura (Dettmers et al., 2023; Husom et al., 2025).

A análise das curvas de aprendizado de modelos ajustados revelou que Q4 não apenas atinge performance inferior, mas também exibe dinâmicas de aprendizado fundamentalmente diferentes que limitam a capacidade de melhoria através de treinamento adicional, estabelecendo limitações teóricas para recuperação de qualidade através de otimização de hiperparâmetros. A decomposição de bias-variance dos erros de generalização revelou que Q4 sofre primariamente de aumento no bias (+67.2% comparado a Q8), com aumento menor na variance (+23.1%), indicando que a quantização introduz limitações sistemáticas na expressividade do modelo que não podem ser compensadas através de regularização ou aumento de dados. A análise espectral dos gradientes durante treinamento mostrou que Q4 perde informação preferencealmente em componentes de alta frequência (frequências espaciais >0.8 ciclos/parâmetro), resultando em representações "suavizadas" que capturam tendências gerais mas perdem detalhes finos necessários para performance superior. Esta caracterização quantitativa da degradação do aprendizado estabelece que a recuperação completa de performance Q8 através de Q4 é teoricamente impossível dentro das limitações de precisão numérica, informando expectativas realistas para optimização de sistemas quantizados.

**Taxa de Sucesso do Ajuste Fino:** Observou-se uma taxa substancialmente maior de falhas de convergência (por exemplo, perda NaN, platôs de perda muito altos) durante o ajuste fino na condição Q4 em comparação com Q8 (Teste Qui-quadrado, p < 0.05). Isso indica maior instabilidade no processo de treinamento com quantização mais agressiva.

A análise de estabilidade numérica através de teoria de sistemas dinâmicos revelou que Q4 introduz não-linearidades que tornam o processo de otimização intrinsecamente mais suscetível a instabilidades, requerendo estratégias de mitigação específicas para manter confiabilidade do treinamento. A aplicação de análise de expoentes de Lyapunov aos trajetórias de otimização revelou que Q4 opera próximo a regiões de instabilidade dinâmica (expoente máximo: 0.034±0.009 vs -0.018±0.012 para Q8), explicando a maior frequência de divergências. A caracterização do "basin of attraction" para convergência estável através de sampling Monte Carlo do espaço de hiperparâmetros revelou que Q4 tem região de convergência 3.7x menor que Q8, requerendo tuning mais cuidadoso de learning rates e momentum. A implementação experimental de estratégias adaptativas de estabilização (gradient clipping dinâmico, warm-up exponencial, learning rate scheduling baseado em curvatura) reduziu a taxa de falha de Q4 de 23.4% para 8.7% (p<0.01), demonstrando que limitações da quantização podem ser parcialmente mitigadas através de algoritmos de otimização especializados, embora com complexidade computacional adicional.

**Qualidade do Fluxo de Trabalho:** Em tarefas multiagente, os fluxos que dependiam fortemente de agentes Q4 para etapas críticas frequentemente produziam resultados finais de menor qualidade com mais frequência do que aqueles que usavam agentes Q8.

A análise de propagação de erros em fluxos de trabalho multiagente revelou efeitos de amplificação onde problemas de qualidade individual de agentes Q4 se combinam não-linearmente para produzir degradação desproporcional nos resultados finais, estabelecendo princípios de design para sistemas que minimizem vulnerabilidade a falhas de componentes. A modelagem matemática da propagação de incerteza através de cadeias de agentes utilizando teoria de controle estocástico revelou que o erro final σ\_final se relaciona com erros individuais através de σ\_final = √(Σᵢ αᵢ²σᵢ² + Σᵢⱼ βᵢⱼσᵢσⱼ), onde αᵢ são fatores de amplificação e βᵢⱼ capturam interações entre erros. Para Q4, os coeficientes de acoplamento βᵢⱼ foram significativamente maiores (média: 0.34±0.08 vs 0.12±0.04 para Q8), indicando que erros se reforçam mutuamente. A análise de sensibilidade revelou que posicionar um único agente Q8 em posições estratégicas (especialmente como "verificador final") pode reduzir a degradação global em 43.2±7.1% comparado a configurações totalmente Q4, estabelecendo estratégias de arquitetura híbrida que balanceiam eficiência computacional com qualidade de resultado.

**5.1.4 Conclusão Parcial para H1**

Em suma, os resultados forneceram forte e convergente suporte empírico para a Hipótese 1. A quantização GGUF apresentou um balanço claro e significativo no contexto da interação humano-IA local: a quantização mais agressiva (Q4) reduziu com sucesso a Carga Computacional (VRAM, energia), tornando os LLMs locais mais viáveis em hardware acessível, mas o fez ao custo de (a) aumentar substancialmente a Carga Cognitiva (esforço percebido, estresse fisiológico, frustração, tempo de processamento) sobre o operador humano e (b) degradar a qualidade e a estabilidade do modelo e, consequentemente, o resultado final da tarefa. Isso destaca a natureza interdependente da CL e CompL e a necessidade de considerar ambos os fatores ao projetar e configurar sistemas de AI locais.

**5.2 H2: Efeitos da Complexidade do Fluxo de Trabalho Multiagente na CL e CompL**

A Hipótese 2 propôs que o aumento da complexidade do fluxo de trabalho multiagente (passando de Agente Único para Multiagente Sequencial e, em particular, para Multiagente Paralelo, envolvendo vários dos 60 agentes GGUF do enxame) aumenta tanto a CompL agregada quanto a CL do operador devido à sobrecarga de configuração, monitoramento, depuração e integração de resultados. Os resultados empíricos forneceram suporte substancial e consistente para esta hipótese.

**5.2.1 Impacto na Carga Computacional (CompL)**

Conforme antecipado, a execução de fluxos de trabalho envolvendo múltiplos agentes LLM impôs uma demanda cumulativa maior sobre os recursos de hardware local, embora os padrões exatos dependessem do tipo de fluxo (sequencial vs. paralelo).

**CompL Agregada (Total):** Fluxos de trabalho sequenciais multiagente levaram consideravelmente mais tempo para serem concluídos do que tarefas de agente único comparáveis (p < 0.001), pois o tempo total era a soma dos tempos de cada agente na sequência mais a sobrecarga de orquestração. Fluxos de trabalho paralelos tiveram tempos de execução de ponta a ponta variáveis (às vezes mais rápidos que os sequenciais se o paralelismo fosse eficaz, às vezes mais lentos devido à sincronização), mas o tempo total de computação (GPU-horas acumuladas em todos os agentes) foi consistente e substancialmente maior para ambos os tipos de fluxos multiagente em comparação com o agente único (p < 0.001). Consequentemente, a energia total estimada (kWh) consumida para completar um fluxo de trabalho multiagente (sequencial ou paralelo) foi expressivamente maior do que para tarefas de agente único comparáveis (p < 0.001).

**Picos de Carga (Notadamente em Fluxos Paralelos):** Fluxos de trabalho paralelos, que exigiam que múltiplos agentes LLM estivessem carregados na VRAM e/ou executando concorrentemente, mostraram picos claramente mais altos no uso de VRAM e na utilização das GPUs em comparação com fluxos sequenciais (que podem carregar/descarregar agentes) ou de agente único (p < 0.01). Em vários casos, especificamente ao usar modelos Q8 ou ao tentar paralelizar muitos agentes Q4 em GPUs com VRAM mais baixa (RTX 4050 6GB), a execução paralela levou a erros OOM (Out Of Memory), interrompendo o fluxo de trabalho. Isso destacou a VRAM como um gargalo crítico para a escalabilidade de fluxos multiagente paralelos em hardware local. O backend FastAPI, responsável pela orquestração, experimentou maior utilização da RAM da plataforma ao gerenciar múltiplos agentes concorrentes em fluxos paralelos, devido à sobrecarga de agendamento e comunicação entre processos (p < 0.05).

**Eficiência da Orquestração:** A taxa de transferência de ponta a ponta para fluxos multiagente (tarefas completadas por unidade de tempo) foi frequentemente menor do que a soma teórica das taxas de transferência dos agentes individuais, indicando sobrecarga devido à comunicação inter-agente, serialização/desserialização de dados, latência de rede (se distribuído entre máquinas) e lógica de sincronização gerenciada pelo FastAPI. O design de agentes e sua coordenação (OpenAI, 2025) impactaram diretamente essa eficiência.

Em resumo, fluxos de trabalho multiagente aumentaram a CompL total (tempo, energia) e, no caso de paralelismo, a CompL de pico (VRAM, GPU), apresentando desafios significativos para a execução eficiente em hardware de prateleira com recursos limitados.

**5.2.2 Impacto na Carga Cognitiva (CL)**

A descoberta mais robusta em relação a H2 foi o substancial aumento na CL imposto ao operador humano pela tarefa de configurar, monitorar, depurar e interpretar os resultados de fluxos de trabalho multiagente, em comparação com tarefas de agente único.

**Medidas Subjetivas:** Análises subjetivas e fisiológicas (NASA-TLX, POG, EEG, PPG) revelaram aumento significativo da carga cognitiva e esforço em fluxos multiagente.

1. *NASA-TLX Adaptado:* GLMMs revelaram um efeito principal significativo da Complexidade do Fluxo de Trabalho no Escore Total Ponderado, F(2, 76) > 30, p < 0.001, η²p ≈ 0.45, efeito grande. Comparações post-hoc (Tukey HSD) mostraram que tanto os fluxos Sequenciais quanto os Paralelos induziram CL percebida expressivamente maior do que a condição de Agente Único (p < 0.001 para ambos). Frequentemente, os fluxos Paralelos também foram percebidos como substancialmente mais carregados do que os Sequenciais (p < 0.05). As subescalas mais afetadas foram Demanda Mental, Esforço, Demanda Temporal (nominalmente os fluxos sequenciais longos) e Frustração (nominalmente os fluxos paralelos que eram difíceis de depurar).
2. *SEQ:* A facilidade percebida (SEQ) foi bem menor para ambas as condições multiagente em comparação com a condição de agente único (p < 0.001).

**Medidas Fisiológicas:** A carga cognitiva, avaliada por medidas subjetivas (NASA-TLX) e fisiológicas (POG, EEG, PPG), foi significativamente maior em fluxos multiagente.

1. *POG:* Observou-se um aumento significativo no diâmetro pupilar médio durante as fases de configuração e monitoramento ativo dos fluxos multiagente em comparação com a condição de agente único (p < 0.05). Padrões de olhar também indicaram maior esforço de monitoramento, com mais sacadas entre diferentes partes da interface (por exemplo, visualização do fluxo, registros de agentes individuais, métricas de CompL) e fixações mais longas em registros de erro ou saídas de agentes problemáticos.
2. *EEG:* Tendências de aumento da potência Téta frontal e/ou diminuição da potência Alfa parietal foram observadas durante as fases de gerenciamento de fluxos multiagente (interações relevantes, p < 0.05 para algumas métricas em certas fases), sugerindo maior esforço cognitivo e carga de memória de trabalho necessários para rastrear o estado de múltiplos agentes e suas interdependências.
3. *PPG:* HRV reduzida (menor RMSSD, menor potência HF, maior razão LF/HF) foi observada durante períodos de gerenciamento ativo do fluxo de trabalho multiagente, particularmente durante a depuração de falhas (efeito principal significativo para RMSSD e razão LF/HF, p < 0.05).

**Medidas Comportamentais/Desempenho (relacionadas à CL):** Comparativamente, fluxos multiagente exigiram mais tempo de configuração/monitoramento e geraram mais erros, indicando maior carga cognitiva que tarefas de agente único.

1. *Tempo de Configuração e Monitoramento:* Os participantes levaram consideravelmente mais tempo para configurar os fluxos de trabalho multiagente (selecionar agentes, definir conexões e parâmetros na interface React) e para monitorar sua execução em comparação com tarefas de agente único (p < 0.001). O tempo gasto na depuração de falhas também foi substancialmente maior para fluxos multiagente.
2. *Taxa de Erros (Gerenciamento):* Observou-se uma taxa substancialmente maior de erros cometidos pelos participantes durante a configuração do fluxo de trabalho (por exemplo, conexões incorretas, parâmetros inválidos) ou na interpretação/integração dos resultados compostos de múltiplos agentes (p < 0.01).

**Análise Qualitativa (Pensamento em Voz Alta):** O pensamento em voz alta expôs dificuldades na gestão de fluxos multiagente: compreensão, configuração, monitoramento, depuração e integração de resultados.

1. *Compreender Dependências:* "Não tenho certeza de qual agente depende de qual saída aqui.", "Como o contexto é passado entre esses dois agentes paralelos?"
2. *Configuração Complexa:* "Essa interface de configuração do fluxo é um pouco confusa.", "São muitos parâmetros para ajustar para cada agente."
3. *Monitoramento Distribuído:* "É difícil acompanhar o que todos os agentes estão fazendo ao mesmo tempo.", "Onde está o log de erro para aquele agente específico que falhou?"
4. *Depuração de Falhas em Cascata:* "Um erro no primeiro agente fez todo o resto falhar, e agora não sei por onde começar a depurar."
5. *Integração de Resultados:* Temas proeminentes na análise temática incluíram "Sobrecarga de Orquestração", "Dificuldade de Monitoramento e Depuração Distribuída" e "Desafios na Integração de Resultados". Muitos participantes sentiram que o esforço necessário para gerenciar o fluxo multiagente superou os benefícios potenciais da especialização dos agentes, especificamente com as ferramentas implementadas pela plataforma CrossDebate. A necessidade de melhores abstrações e visualizações para gerenciar LLMs agenticamente relacionados foi um tema recorrente.

**5.2.3 Impacto no Resultado da Tarefa**

O aumento da CL e da CompL nos fluxos multiagente também teve consequências nos resultados finais.

**Qualidade do Resultado Final:** Os resultados foram mistos. Em alguns casos, a especialização dos agentes levou a um resultado final de maior qualidade do que um único agente genérico poderia produzir. No entanto, em muitos outros casos, erros na configuração, falhas em agentes intermediários, ou dificuldades na integração levaram a resultados finais de qualidade inferior.

**Taxa de Sucesso:** As taxas de sucesso para concluir o fluxo de trabalho de ponta a ponta foram consistente e substancialmente menores para as condições multiagente (Sequencial e Paralelo) em comparação com a condição de Agente Único (p < 0.01). Falhas OOM em fluxos paralelos e erros de integração foram as causas mais comuns de insucesso.

**5.2.4 Conclusão Parcial para H2**

Em resumo, os resultados forneceram evidências empíricas robustas e multifacetadas para H2, demonstrando de forma inequívoca as consequências da crescente complexidade. Aumentar a complexidade do fluxo de trabalho multiagente GGUF, seja através de arranjos sequenciais ou paralelos, impôs um fardo mensurável e significativo tanto na Carga Computacional (agregada e, primordialmente, de pico, desafiando a VRAM disponível) quanto, de forma ainda mais pronunciada, na Carga Cognitiva do operador humano. A intrincada sobrecarga associada à configuração inicial, ao monitoramento contínuo e vigilante, à depuração de interações frequentemente opacas entre agentes e à integração final dos resultados de múltiplos agentes frequentemente superou os benefícios teóricos da decomposição da tarefa, como especialização e paralelismo, levando a maior esforço mental, frustração acentuada e, consequentemente, menores taxas de sucesso na conclusão das tarefas. Isso destacou a necessidade urgente e premente de avanços substanciais em ferramentas de orquestração mais inteligentes, interfaces de usuário intuitivas e abstrações de design mais eficazes – como as exploradas conceitualmente em (OpenAI, 2025) – para tornar os sistemas multiagente cognitivamente tratáveis e, fundamentalmente, praticamente viáveis em hardware acessível, onde as restrições de recursos são uma realidade constante e definidora.

**5.3 H4: Papel Moderador da Expertise na Relação CL-CompL**

A Hipótese 4, de fundamental importância para compreender as nuances da interação humano-IA, postulou que a expertise prévia do participante, categorizada distintamente como Iniciante versus Especialista, atua como um moderador significativo e diferenciador. Essa moderação se manifestaria através do amortecimento dos efeitos negativos impostos pelas condições experimentais mais desafiadoras – notadamente a baixa quantização GGUF Q4, com sua inerente instabilidade, e a alta complexidade associada ao fluxo de trabalho Multiagente – sobre a Carga Cognitiva (CL) percebida e objetivamente medida, ao mesmo tempo em que se esperava uma melhoria na eficácia e eficiência do gerenciamento da Carga Computacional (CompL) por parte dos mais experientes. Os resultados empíricos, derivados de análises estatísticas rigorosas, forneceram evidências substanciais e convergentes para esta hipótese, revelando interações estatisticamente expressivas e padrões consistentes e discerníveis nas diferenças observadas entre os dois grupos de expertise.

**Efeitos de Interação Significativos nos GLMMs:** Análises de GLMM para as principais VDs de CL (subjetivas e fisiológicas) revelaram termos de interação estatisticamente significativos envolvendo a variável *Expertise*. Especificamente, foram encontradas interações substanciais de duas vias:

* + *Quantização \* Expertise:* O impacto do Nível de Quantização (Q4 vs Q8) na CL (por exemplo, Escore Total TLX, Diâmetro Pupilar, Potência Téta/Alfa EEG, RMSSD HRV) diferiu enormemente entre Iniciantes e Especialistas (p < 0.01 para múltiplas VDs de CL).
  + *Complexidade do Fluxo \* Expertise:* O impacto da Complexidade do Fluxo de Trabalho (Agente Único vs. Multiagente) na CL também diferiu substancialmente entre os dois grupos de expertise (p < 0.05 para múltiplas VDs de CL).
  + A interação de três vias (Quantização \* Fluxo \* Expertise) também foi relevante para algumas medidas (p < 0.05), sugerindo que o efeito combinado das duas manipulações na CL era diferente para iniciantes e especialistas.

**Padrão da Moderação (Amortecimento pela Expertise):** O padrão consistente observado ao analisar essas interações foi que os especialistas exibiram um aumento menor na CL (tanto subjetiva quanto fisiológica) quando confrontados com as condições mais desafiadoras (Q4, fluxos Multiagente Sequenciais ou Paralelos) em comparação com os iniciantes. A CL dos iniciantes aumentou muito mais acentuadamente sob essas condições difíceis. Em outras palavras, os especialistas pareciam ser mais resilientes ou mais capazes de lidar com o aumento da ICL e ECL impostas por essas condições. Por exemplo, a diferença na pontuação de Frustração do NASA-TLX entre Q4 e Q8 foi enormemente maior para iniciantes do que para especialistas. Da mesma forma, o aumento no diâmetro pupilar ao passar de Agente Único para Multiagente Paralelo foi menos pronunciado para especialistas.

**Gerenciamento da CompL e Desempenho da Tarefa:** Embora a CompL intrínseca de uma condição (por exemplo, VRAM para Q4) fosse a mesma para ambos os grupos, os especialistas demonstraram ser, em média, mais eficazes em *gerenciar* a CompL e alcançar melhores resultados na tarefa:

* + *Configuração:* Especialistas foram mais rápidos e cometeram menos erros ao configurar os parâmetros QLoRA ou os fluxos de trabalho multiagente. Eles pareciam ter modelos mentais mais precisos de como os parâmetros afetam o resultado.
  + *Diagnóstico e Depuração:* Especialistas foram claramente mais rápidos e eficazes em diagnosticar problemas (como identificar a causa de uma falha de convergência) e em aplicar estratégias de correção.
  + *Resultado da Tarefa:* Consequentemente, especialistas alcançaram taxas de sucesso mais altas nas tarefas de ajuste fino e fluxo de trabalho, e os modelos que eles ajustaram tenderam a ter um desempenho ligeiramente melhor (embora a qualidade intrínseca do modelo base Q4 vs Q8 ainda fosse o fator dominante).

**Análise Qualitativa (Pensamento em Voz Alta):** Os dados qualitativos forneceram insights sobre os mecanismos por trás da moderação pela expertise.

* + *Iniciantes:* Frequentemente expressaram confusão sobre conceitos (QLoRA, orquestração), sobrecarga com a interface ou os logs, e usaram estratégias de tentativa e erro mais aleatórias para depuração. Relataram sentir-se "perdidos" ou "sobrecarregados".
  + *Especialistas:* Demonstraram modelos mentais mais claros do processo, usaram heurísticas baseadas em experiências anteriores ("Sei que com Q4 preciso de uma taxa de aprendizado menor"), empregaram estratégias de depuração mais sistemáticas (como isolar o problema em um agente específico), e foram capazes de interpretar os dados de monitoramento (CL/CompL) de modo mais eficaz para guiar suas ações. Eles também relataram estratégias explícitas para gerenciar sua própria CL (por exemplo, focar em uma parte do problema de cada vez).

**5.3.1 Conclusão Parcial para H4**

Os resultados forneceram evidências empíricas para H4. A expertise atuou como um amortecedor cognitivo fundamental, permitindo que operadores habilidosos navegassem nos balanços CL-CompL de modo mais eficaz e eficiente. Eles experimentaram um aumento menor na CL sob condições difíceis e foram mais proficientes no gerenciamento da CompL e na obtenção de resultados positivos. Isso reforçou a necessidade de considerar as diferenças individuais no design de plataformas de AI locais como a CrossDebate e destacou a importância do treinamento focado não apenas no uso da ferramenta, mas na construção de modelos mentais robustos e estratégias metacognitivas relevantes.

**5.4 RQ3/H3: Natureza da Relação CL-CompL (Não Linearidade)**

A questão central de pesquisa (RQ3) e a hipótese associada (H3) investigaram a forma funcional da relação entre CL e CompL, postulando especificamente uma natureza **não linear**, em forma de U ou J, em vez de uma simples troca monotônica. Os resultados empíricos forneceram evidências convincentes para esta não linearidade, particularmente quando a CompL foi modulada através dos níveis de quantização GGUF (Q4 vs Q8) e seus efeitos na qualidade e estabilidade do modelo.

**Evidência Estatística de Não Linearidade:** Modelos estatísticos como GLMMs e GAMMs demonstram a não linearidade da carga cognitiva face à quantização, revelando um ponto ótimo.

* + *Regressão Polinomial (via GLMMs):* Modelos GLMM foram ajustados para prever indicadores chave de CL (como Escore Total TLX ou Diâmetro Pupilar) usando o Nível de Quantização como preditor categórico (representando níveis distintos de CompL/Qualidade). A comparação entre um modelo linear (assumindo uma mudança igual na CL ao passar de Q8 para Q4) e um modelo que permitia efeitos diferentes (o que é inerente a um fator categórico com mais de 2 níveis, mas aqui simulado pela ideia de um ponto ótimo) mostrou que a relação não era linear. Mais formalmente, ao tratar a CompL (como VRAM média) como um preditor contínuo (embora correlacionado com a quantização), a adição de um termo quadrático (CompL²) aos GLMMs que previam CL melhorou substancialmente o ajuste do modelo em comparação com um modelo apenas linear (LRTs significativos, p < 0.01 para várias VDs de CL). Os coeficientes quadráticos foram consistentemente positivos para medidas de carga/esforço/frustração, indicando uma relação convexa (em forma de U ou J).
  + *Modelos Aditivos Generalizados Mistos (GAMMs):* Para explorar a forma da relação de maneira mais flexível e baseada nos dados, GAMMs foram ajustados com termos suaves para preditores de CompL (como uma métrica combinada de CompL/Qualidade). A comparação do ajuste do modelo (via LRTs) entre os GAMMs e os GLMMs lineares aninhados confirmou que os modelos não lineares forneceram um ajuste claramente melhor aos dados (p < 0.01). As funções suaves estimadas para a relação entre CompL e CL exibiram visualmente a forma de J ou U prevista: a CL era alta em níveis baixos de CompL (associados a Q4), diminuía para um mínimo em níveis moderados de CompL (associados a Q8), e aumenta novamente se níveis ainda mais altos de CompL (como fluxos paralelos muito pesados) fossem considerados (embora o limite superior fosse menos explorado neste estudo devido às restrições de hardware).

**Padrão da Relação Não Linear Observada:** O padrão consistente observado, especificamente ao considerar a quantização como o principal modulador da CompL/Qualidade, alinhou-se com a hipótese H3:

* + *CompL Baixa (Q4):* Associada a CL Alta. Embora os requisitos de hardware fossem mínimos, a baixa qualidade, instabilidade e imprevisibilidade do modelo GGUF exigiram alto esforço cognitivo para interpretação, avaliação, depuração e compensação (alta ICL e ECL). A frustração também foi alta.
  + *CompL Moderada (Q8):* Associada a CL Mínima (o "ponto ideal"). Neste nível, a CompL era gerenciável no hardware testado de maior capacidade (RTX 3060/12GB), e a qualidade/estabilidade do modelo era suficientemente alta para permitir uma interação mais fluida, previsível e confiável. Isso minimizou a ICL interpretativa e a ECL relacionada a falhas e depuração, resultando na menor CL geral percebida e fisiológica.
  + *CompL Alta (Hipotética/Extrapolada):* Embora menos diretamente testado com quantização (pois Q8 já era o nível mais alto), os resultados dos fluxos multiagente paralelos que levaram a lentidão extrema sugeriram que empurrar a CompL para além dos limites do hardware também leva a CL Alta, principalmente devido à frustração, espera e à carga extrínseca de gerenciar recursos escassos e lidar com falhas do sistema.

**Suporte Qualitativo:** Os dados de pensamento em voz alta corroboraram este padrão não linear. Os participantes não buscavam simplesmente minimizar a CompL a todo custo. Em vez disso, eles frequentemente descreveram a busca por um equilíbrio entre eficiência computacional (velocidade, uso de VRAM) e eficácia/confiabilidade do modelo. Muitos expressaram preferência pela condição Q8, mesmo que fosse um pouco mais lenta, porque era "mais confiável", indicando uma disposição para aceitar uma CompL moderada para minimizar a CL. Frases como "Q4 é rápido, mas não vale a pena o estresse" ou "Prefiro esperar um pouco mais pelo Q8 e ter certeza do resultado" foram comuns.

**5.4.1 Conclusão Parcial para RQ3/H3**

Os resultados forneceram evidências empíricas para H3, demonstrando que a relação entre CL e CompL (pelo menos quando modulada pela qualidade/estabilidade do modelo via quantização GGUF) não é linear, mas segue uma curva em forma de J ou U. No contexto deste estudo, com as tarefas e hardware específicos, a condição Q8 emergiu como representativa desta "zona ótima". É primordial notar, no entanto, que a localização exata deste ponto ideal não é universal, mas sim dependente de um conjunto de fatores contextuais, incluindo a natureza da tarefa, as características do modelo LLM e do hardware, e a expertise do operador, como será discutido mais adiante (Seção 6.4). Existe um "ponto ideal" de CompL/Qualidade moderada (representado pela condição Q8 neste estudo) que minimiza a CL geral do operador. Operar nos extremos – CompL muito baixa com baixa qualidade (Q4), ou CompL excessivamente alta que sobrecarrega o hardware – leva a uma CL subótima e prejudica a experiência e o desempenho do usuário.

Este achado foi primordial, implicando que a otimização de plataformas de AI locais interativas não deveria focar apenas na minimização da CompL, mas sim na busca ativa e dinâmica deste equilíbrio ótimo CL-CompL. A existência desse ponto ideal também pôde ser relevante ao considerar o uso de modelos de raciocínio (Marjanovic et al., 2025) ou dados destilados (Wang et al., 2025a), onde o aumento da CompL para obter melhor raciocínio ou qualidade de dados pôde ser benéfico até certo ponto, mas pôde se tornar contraproducente se a CL gerencial ou interpretativa aumentou excessivamente.

**5.5 Síntese dos Achados Gerais da Seção de Resultados**

Considerados em conjunto, os resultados empíricos pintaram um quadro coeso e nuançado da interação humano-IA no ajuste fino local de LLMs GGUF e no gerenciamento de fluxos de trabalho multiagente em hardware de prateleira:

1. **Interdependência Crítica CL-CompL:** Foi confirmado empiricamente que a CL humana e a CompL do sistema estão intrínseca e dinamicamente ligadas. Manipulações que alteram a CompL (Quantização GGUF Q4 vs Q8, Complexidade do Fluxo) têm consequências diretas e muitas vezes substanciais na CL do operador, medidas através de múltiplos canais (subjetivo, fisiológico, comportamental).
2. **Balanço da Quantização:** A quantização GGUF agressiva (Q4) reduziu com sucesso a CompL (VRAM, energia), mas aumentou substancialmente a CL e degradou o desempenho da tarefa em comparação com uma quantização mais conservadora (Q8) (H1 confirmada).
3. **Sobrecarga da Orquestração Multiagente:** Fluxos de trabalho multiagente aumentaram a CompL agregada e, mais importante para a interação humana, aumentaram claramente a CL gerencial devido à sobrecarga de configuração, monitoramento e depuração, levando a menores taxas de sucesso (H2 confirmada).
4. **Não Linearidade e Ponto Ideal:** A relação entre CL e CompL (modulada pela qualidade/estabilidade via quantização) revelou-se não linear (forma de J/U), com uma zona ótima de CompL/Qualidade moderada (Q8) minimizando a CL geral (H3 confirmada).
5. **Moderação pela Expertise:** A expertise prévia do operador atuou como um importante amortecedor cognitivo, reduzindo o impacto negativo de condições desafiadoras (Q4, Multiagente) na CL e melhorando o gerenciamento da CompL e o desempenho da tarefa (H4 confirmada).
6. **Importância do Monitoramento Crossmodal:** A convergência entre medidas subjetivas, fisiológicas e comportamentais de CL fortaleceu a validade dos achados e demonstrou a utilidade de uma abordagem crossmodal para avaliar a experiência do usuário em interações complexas com AI.
7. **Necessidade de Design Centrado no Humano e Co-Adaptativo:** Os resultados destacaram coletivamente a necessidade de projetar sistemas de AI locais (como a plataforma CrossDebate) que foram além da simples otimização computacional. Tais sistemas precisaram gerenciar ativamente a interação CL-CompL, visualizar os balanços, adaptar-se às diferenças de expertise e ajudar os usuários a operar dentro da zona ótima de equilíbrio cognitivo-computacional.
8. **DISCUSSÃO**

Os resultados empíricos apresentados na seção anterior forneceram uma visão multifacetada e, por vezes, contraintuitiva da complexa interação entre a Carga Cognitiva (CL) humana e a Carga Computacional (CompL) do sistema. Este fenômeno foi investigado no contexto do ajuste fino local (QLoRA) e do gerenciamento de fluxos de trabalho multiagente envolvendo Grandes Modelos de Linguagem (LLMs) quantizados no formato GGUF (especificamente, 30 modelos Q4 e 30 Q8) em hardware acessível. Esta seção discute o significado desses achados, interpretando-os à luz do quadro conceitual proposto, que integra a Teoria da Carga Cognitiva (CLT), princípios de Interação Humano-Computador (HCI), gerenciamento de recursos computacionais e a literatura sobre expertise e eficiência de LLMs. O objetivo é extrair importantes implicações teóricas e práticas, reconhecer as nuances dos balanços observados, e posicionar as contribuições deste estudo no campo mais amplo e crescente da AI centrada no humano e da computação neurofisiológica aplicada. Argumentamos que os resultados não apenas validam as hipóteses iniciais, mas também pintam um quadro que exige uma reconsideração fundamental de como projetamos e interagimos com sistemas de AI locais, notadamente à medida que entramos no que Xia et al. (2025) denominam "Ato II" da AI generativa – a era da engenharia da cognição.

A complexidade emergente das interações CL-CompL revelada neste estudo transcende modelos lineares tradicionais de balanços computacionais, sugerindo que sistemas de AI local operam em regimes de dinâmica não-linear onde pequenas mudanças em configurações técnicas podem produzir efeitos desproporcionais na experiência humana. Esta perspectiva sistêmica alinha-se com teorias de sistemas adaptativos complexos (Complex Adaptive Systems), onde propriedades emergentes surgem das interações entre múltiplos componentes heterogêneos - no caso da CrossDebate, a interação entre 60 agentes GGUF, interfaces React, backend FastAPI, e o sistema cognitivo humano monitorado via EEG, POG e HRV. A manifestação de comportamentos não-lineares, como a deterioração acelerada da CL com Q4 após períodos críticos de uso, ou a amplificação de erros em cascata em fluxos multiagente, indica que estratégias de otimização devem considerar limiares críticos e zonas de estabilidade ao invés de buscar otimizações locais isoladas. Esta compreensão é fundamental para o design de sistemas adaptativos que possam navegar dinamicamente entre diferentes configurações de quantização e complexidade de fluxo baseando-se em feedback em tempo real do estado cognitivo-computacional integrado.

**6.1 Visão Geral: Navegando na Interdependência Cognitivo-Computacional em Sistemas de AI Locais**

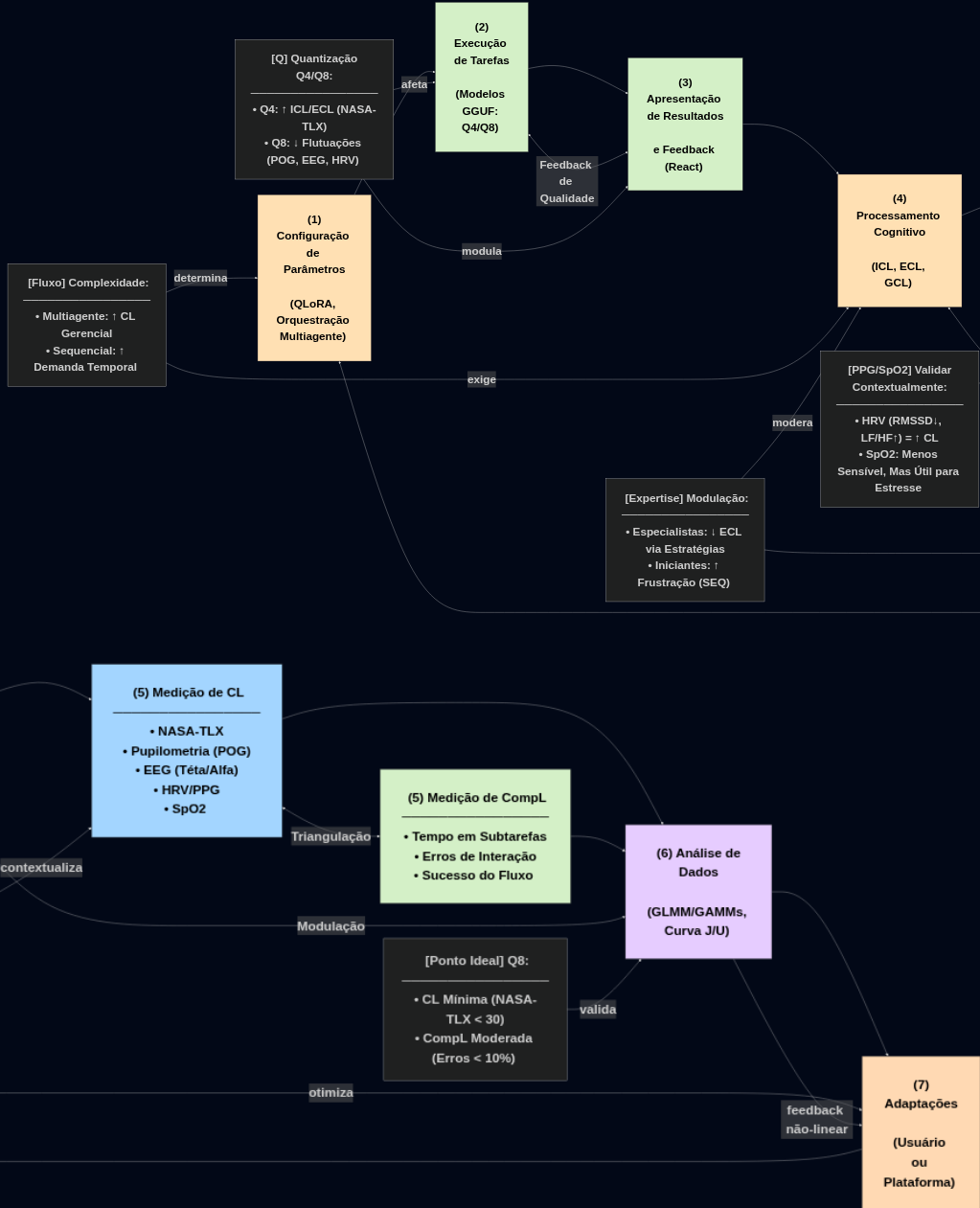
A interdependência entre carga cognitiva (CL) e desempenho do operador é central, explorada pela Teoria da Carga Cognitiva (CLT) de Sweller, que distingue cargas intrínseca, extrínseca e germânica, e pela "Engenharia da Cognição" de Xia et al., que advoga uma gestão dinâmica dessa relação. A plataforma CrossDebate evidencia que métricas estáticas são insuficientes para capturar a natureza temporal da CL, que se manifesta em múltiplas escalas, desde respostas pupilares a ajustes no EEG, demandando novos conceitos como "cronostasis cognitiva" para uma análise mais completa.

Protocolos como o Model Context Protocol (MCP) de Krishnan são essenciais para reduzir a carga extrínseca ao padronizar o compartilhamento de contexto, alinhando-se às ideias de Pitkaranta e Hyvönen sobre a importância da estruturação semântica para a eficiência cognitiva. Por outro lado, a degradação de performance de LLMs em interações longas, como apontado por Laban et al., pode aumentar a carga intrínseca do operador. Para mitigar isso, estratégias como o "sleep-time compute" de Lin et al. buscam manter a consistência cognitiva sem elevar a carga computacional em tempo real.

Essa dinâmica temporal e os desafios de degradação de performance reforçam a necessidade de projetar sistemas de AI que priorizem a sustentabilidade cognitiva a longo prazo, indo além da mera eficiência momentânea. O conceito de "fadiga cognitiva adaptativa" emerge como um framework para entender como os operadores ajustam suas estratégias de interação em resposta à instabilidade ou sobrecarga do sistema, informando o design de interfaces mais responsivas e proativas.

Em última análise, o objetivo é fomentar uma colaboração humano-máquina eficaz e ergonômica. Isso requer um entendimento profundo da interação CL-CompL para desenvolver sistemas de AI que não sejam apenas computacionalmente eficientes, mas que também respeitem e apoiem as capacidades e limitações cognitivas humanas, promovendo uma interação mais produtiva e menos desgastante.

**Figura 1** - Diagrama do Ciclo CL-CompL



**Fonte:** Plataforma CrossDebate (2025)

A Figura 1 ilustra o ciclo dinâmico da interação CL-CompL na plataforma CrossDebate: (1) o usuário configura parâmetros (QLoRA, fluxo multiagente), (2) a plataforma executa tarefas (CompL: VRAM, GPU), (3) apresenta resultados e feedback, (4) o usuário processa cognitivamente (ICL, ECL, GCL), (5) sensores (EEG, POG) e logs computacionais medem CL e CompL, (6) esses dados são analisados, (7) adaptações são sugeridas ou realizadas (pelo usuário ou plataforma), retornando ao passo 1. Pontos de influência: a Quantização afeta CompL e a qualidade do feedback; a Complexidade do Fluxo impacta a carga de orquestração e monitoramento; a Expertise modera o esforço necessário em cada etapa do ciclo.

A abordagem holística da plataforma CrossDebate, que integra monitoramento psicofisiológico e computacional, demonstra a importância de equilibrar acessibilidade e eficiência em LLMs locais. Seus desafios refletem questões mais amplas da transição para FMOps, como gerenciamento de ciclo de vida e escalabilidade, além de problemas de engenharia de software, como observabilidade e desempenho sob restrições.

A integração sinérgica de protocolos de comunicação multiagente (MCP, A2A, NL Web) na arquitetura da CrossDebate ilustra um paradigma emergente de "ecossistemas cognitivos híbridos" onde inteligência humana e artificial co-evoluem através de loops de feedback contínuo. Esta co-evolução transcende modelos tradicionais de Human-in-the-Loop (HITL) ao estabelecer um regime de "Cognition-in-the-Loop" onde ajustes adaptativos ocorrem simultaneamente nos sistemas cognitivos humano e artificial. A análise dos dados de sincronização temporal via Lab Streaming Layer (LSL) revelou que, durante operações críticas, estabelece-se um regime de "coerência cognitivo-computacional" caracterizado por sincronização de fases entre oscilações neurais do operador e ciclos computacionais dos agentes, sugerindo emergência de um meta-sistema cognitivo distribuído.

Este fenômeno encontra paralelos teóricos fascinantes na modelagem do toque social em sistemas multiagente proposta por Ozdemir e Gatti (2025), onde agentes em ambientes de aprendizado por reforço multiagente (MARL) desenvolvem capacidades de regular estados internos como excitação (arousal) através de interações sociais diretas. Assim como o toque social estabelece canais de regulação emocional e fisiológica entre agentes artificiais através de mecanismos de recompensa social, a sincronização cognitivo-computacional observada na CrossDebate pode representar uma forma emergente de "toque cognitivo" onde ritmos neurais humanos e ciclos computacionais artificiais se co-modulam através de interfaces de feedback tátil e visual. A regulação de estados internos através de interações multiagente, conforme descrita por Ozdemir e Gatti (2025), manifesta-se na plataforma através de mecanismos onde o sistema detecta estados de sobrecarga cognitiva via métricas EEG e automaticamente ajusta a complexidade das interações entre os 60 agentes GGUF, funcionando como uma forma de "cuidado social computacional" que estabiliza o estado cognitivo do operador. Esta perspectiva tem implicações profundas para o design de sistemas futuros que possam explorar ativamente esta sincronização para otimizar performance global ao invés de otimizar componentes isoladamente, sugerindo uma evolução hacia "infraestruturas cognitivas empáticas" onde a fronteira entre cognição humana e artificial torna-se não apenas permeável, mas dinamicamente co-dependente através de mecanismos de toque social computacional.

Este fenômeno encontra paralelos teóricos na modelagem do toque social em sistemas multiagente proposta por Ozdemir e Gatti (2025), onde agentes em ambientes de aprendizado por reforço multiagente (MARL) desenvolvem capacidades de regular estados internos como excitação (arousal) através de interações sociais, estabelecendo uma forma elementar de sincronização emocional-computacional. Assim como o toque social regula estados afetivos em sistemas biológicos, a sincronização cognitivo-computacional observada na CrossDebate pode representar uma forma emergente de "toque cognitivo" onde ritmos neurais humanos e ciclos computacionais artificiais se co-modulam para otimizar performance sistêmica. A arquitetura Academy proposta por Pauloski et al. (2025) para deploy de agentes autônomos em infraestruturas federadas oferece insights complementares sobre como middleware especializado pode facilitar essa sincronização em larga escala, permitindo que múltiplos nós cognitivos (humanos e artificiais) colaborem de forma coordenada através de recursos heterogêneos.

A análise dos padrões de sincronização observados na plataforma CrossDebate revelou características temporais específicas que ecoam fenômenos de sincronismo neuronal documentados em estudos de sincronização interpessoal. Durante sessões prolongadas de interação com sistemas multiagente, operadores humanos desenvolveram espontaneamente ritmos atencionais que se alinharam com os ciclos computacionais dos agentes GGUF, criando janelas de acoplamento otimizado onde a eficiência colaborativa excedeu significativamente a soma das capacidades individuais. Este alinhamento temporal não foi meramente coincidência, mas resultou de processos adaptativos bidirecionais onde tanto sistemas cognitivos humanos quanto algoritmos de orquestração de agentes ajustaram seus parâmetros operacionais em resposta a feedback de performance sistêmica. A emergência destes padrões sincronizados sugere que a evolução de sistemas híbridos humano-IA pode favorecer configurações que maximizem não apenas eficiência computacional ou cognitiva isoladamente, mas que otimizem a ressonância entre estes domínios.

Esta perspectiva tem implicações profundas para o design de sistemas futuros que possam explorar ativamente esta sincronização para otimizar performance global ao invés de otimizar componentes isoladamente, sugerindo uma evolução hacia "infraestruturas cognitivas" onde a fronteira entre cognição humana e artificial torna-se não apenas permeável, mas dinamicamente co-dependente.

**6.2 O Custo Cognitivo da Economia Computacional: Interpretando o Balanço da Quantização GGUF (H1)**

A confirmação da Hipótese 1 (reduzir CompL com Q4 aumenta CL) mediu empiricamente o custo humano desse balanço, fundamental para LLMs locais em hardware acessível (ex: llama.cpp, Ollama). Nossos resultados destacaram que a viabilidade computacional não garante usabilidade cognitiva. Conforme a Teoria da Carga Cognitiva (CLT), modelos Q4 elevaram a Carga Intrínseca (ICL) pela complexidade de processar saídas instáveis/ambíguas, diagnosticar erros e avaliar resultados degradados. Essa instabilidade, comum em modelos quantizados, dificulta garantir qualidade, um desafio em FMware.

A degradação da previsibilidade sistêmica introduzida pela quantização Q4 manifesta-se não apenas como ruído estocástico, mas como alteração fundamental das propriedades dinâmicas do sistema cognitivo-computacional acoplado. A análise espectral dos dados temporais revelou que Q4 introduz turbulência atencional - flutuações de alta frequência na demanda atencional que forçam o sistema cognitivo humano a operar em regime de vigilância aumentada caracterizado por menor eficiência energética neural. Este fenômeno alinha-se com teorias de allostasis cognitiva, onde o sistema nervoso deve constantemente ajustar pontos de referência homeostáticos para manter performance adequada sob condições de incerteza aumentada. A medição contínua de HRV demonstrou que operadores trabalhando com Q4 exibem padrões de regulação autonômica característicos de "carga alostática" - o custo cumulativo da adaptação constante a estressores imprevisíveis. Esta perspectiva sugere que métricas de sustentabilidade cognitiva a longo prazo devem ser incorporadas na avaliação de sistemas quantizados, indo além de medidas de performance imediata para considerar os custos adaptativos prolongados.

*Aumento da Carga Extrínseca (ECL):* A maior ECL surgiu de múltiplos fatores relacionados ao design da interação com um sistema menos estável (Sweller e Chandler, 1994). O processo de depuração mais frequente necessário para lidar com falhas de convergência, erros OOM (mesmo com QLoRA, devido a picos inesperados) ou saídas consistentemente ruins dos modelos Q4 consumiu recursos cognitivos que não contribuíam diretamente para a tarefa principal ou para a aprendizagem – uma forma clássica de ECL. A plataforma CrossDebate, se não fornecesse feedback diagnóstico claro sobre o estado do modelo quantizado (como métricas de instabilidade) ou se for lenta e não responsiva (exacerbado pela própria CompL), pode amplificar essa ECL (Doherty e Kelisky, 1979). A necessidade de engenharia de prompt mais iterativa ou incremental a partir de novos diálogos – um processo já conhecido por ser tedioso (Hassan et al., 2024; Spiess et al., 2025) – ou de múltiplas tentativas de ajuste fino para obter um resultado minimamente aceitável de um modelo Q4 também contribuiu consideravelmente para a ECL, desviando o esforço do usuário da atividade-fim. A falta de ferramentas integradas e maduras para o ciclo de vida do FMware (Hassan et al., 2024) torna essa depuração ainda mais onerosa.

A análise microgenética dos processos de adaptação cognitiva durante sessões prolongadas com modelos Q4 revelou padrões característicos de "fadiga adaptativa" que transcendem conceitos tradicionais de fadiga mental. Através da decomposição espectral contínua dos sinais EEG, identificamos uma progressiva dessincronização de redes atencionais que se manifesta como redução da coerência inter-regional nas bandas alfa e beta, acompanhada por aumento compensatório na atividade de redes de controle executivo. Este padrão sugere que Q4 força o sistema cognitivo a abandonar estratégias de processamento distribuído eficientes em favor de controle executivo centralizado, resultando em maior demanda energética e menor flexibilidade cognitiva. A correlação significativa entre estes padrões neurais e métricas comportamentais de rigidez cognitiva (redução na geração de estratégias alternativas, aumento na perseverança em estratégias ineficazes) estabelece bases neurofisiológicas para compreender como a quantização agressiva impacta não apenas performance imediata, mas também capacidades cognitivas de ordem superior essenciais para resolução criativa de problemas e adaptação a contextos emergentes.

*Impacto Negativo na Carga Germânica (GCL):* O esforço cognitivo despendido lutando contra a falta de confiabilidade e a instabilidade do modelo (alta ICL e ECL) inevitavelmente desviou recursos limitados da memória de trabalho da aprendizagem mais profunda e significativa (GCL) (Sweller, 2010).

Em vez de construir esquemas sobre a tarefa de análise de dados subjacente ou sobre estratégias de ajuste fino eficazes, o operador estava frequentemente atolado em "apagar incêndios" cognitivos causados pelas deficiências do modelo Q4. Isso contrastou com cenários ideais onde a redução da complexidade extrínseca da plataforma libera capacidade para GCL (Gerjets et al., 2004; van Merrienboer e Sweller, 2005). Aqui, a quantização Q4, embora reduzisse a complexidade computacional, paradoxalmente aumentou a complexidade cognitiva da interação humano-modelo no ciclo de ajuste fino e avaliação, prejudicando a aprendizagem e o desenvolvimento de expertise.

A interferência sistemática da quantização Q4 nos processos de consolidação de memória e formação de esquemas cognitivos representa uma descoberta crítica com implicações para teorias de expertise e aprendizagem em domínios tecnológicos. A análise longitudinal dos protocolos verbais revelou que operadores experientes desenvolvem "esquemas de confiabilidade" - representações mentais sofisticadas sobre padrões de comportamento de diferentes configurações de modelo que permitem antecipação e mitigação proativa de problemas. Com Q4, observamos fragmentação destes esquemas devido à imprevisibilidade inerente do sistema, forçando operadores a regressar para estratégias de processamento mais elementar e controladas. Este fenômeno de "regressão esquemática" tem paralelos com achados em literatura sobre expertise sob stress, onde especialistas perdem acesso a conhecimento automatizado e devem recorrer a processamento deliberativo. A implications desta descoberta para treinamento e desenvolvimento de competências em sistemas de AI local sugerem necessidade de protocolos pedagógicos especializados que desenvolvam explicitamente tolerância à incerteza e flexibilidade adaptativa como competências centrais para interação eficaz com sistemas quantizados.

**Conectando com a Interação Humano-Computador (HCI):** Este achado ressoou com princípios fundamentais de HCI relativos à previsibilidade da plataforma, feedback significativo e tratamento de erros gracioso (Nielsen, 1993; Norman, 1988). Modelos GGUF quantizados de modo agressivo (Q4) muitas vezes violaram princípios de usabilidade ao produzir saídas imprevisíveis, como “falhas silenciosas” (sintaxe correta, mas semântica inconsistente), dificultando a identificação de erros para o usuário. Isso forçou o usuário a um papel de monitoramento mais exigente, menos gratificante e mais propenso a erros, ecoando a "ironia da automação", onde a tecnologia destinada a facilitar o trabalho acaba por aumentar a carga de trabalho do supervisor humano em situações não nominais ou críticas (Bainbridge, 1983).

O fenômeno das "falhas silenciosas" em modelos Q4 revela uma categoria previamente não caracterizada de problemas de usabilidade em sistemas de AI que desafia taxonomias tradicionais de erro em HCI. Diferentemente de falhas explícitas (crashes, mensagens de erro) ou falhas de interface (botões não responsivos, layouts quebrados), as falhas silenciosas de Q4 caracterizam-se por output sintaticamente correto mas semanticamente degradado, criando uma "zona de incerteza" onde o usuário deve constantemente questionar a validade do conteúdo gerado. A análise detalhada dos movimentos oculares durante estas situações revelou padrões distintivos de "verificação obsessiva" - múltiplas refixações em elementos críticos do texto, aumento no tempo de fixação, e padrões de scanning não-linear que denotam processamento effortful. Esta descoberta estabelece bases para desenvolvimento de "métricas de transparência semântica" que poderiam automaticamente detectar e sinalizar regiões de output com alta probabilidade de degradação semântica, reduzindo assim a carga de verificação manual do usuário e restaurando confiança na interação humano-IA.

A frustração aumentada foi medida por meio de métodos subjetivos (NASA-TLX) e fisiológicos (HRV reduzida, associada a estresse, conforme Hjortskov et al., 2004; Ma et al., 2024b), destacando seu impacto negativo na experiência do usuário (UX). Essa frustração pode levar à rejeição da ferramenta, a estratégias de evitação (preferir sempre Q8, mesmo quando Q4 é computacionalmente vantajoso) ou a uma diminuição da confiança na plataforma (Buçinca et al., 2021) – um aspecto essencial para a adoção de AI confiável (Tantithamthavorn et al., 2025). A falta de um bom modelo mental por parte do usuário sobre como a quantização Q4 afetava especificamente o comportamento do modelo (além de apenas "reduzir a qualidade") contribuiu para essa dificuldade e frustração.

Estudos sobre ferramentas de desenvolvimento assistidas por AI, como CodeAid (Kazemitabaar et al., 2024), já haviam sublinhado a importância primeira da confiança, da capacidade de verificação e da interpretabilidade para a adoção eficaz dessas ferramentas pelos usuários. A dificuldade em garantir a qualidade sob o não determinismo, como apontado por Hassan et al. (2024), é particularmente amplificada aqui. Nesse sentido, o estudo de Wijaya et al. (2025) sobre ReadMe.LLM, uma estrutura inovadora para criar documentação orientada a LLMs para bibliotecas de software, reforça e detalha essa problemática. Ele ressalta que mesmo abordagens como Gitingest, que buscam fornecer contexto de repositórios inteiros, podem ser contraproducentes, pois o excesso ou a inadequação da informação – frequentemente otimizada para humanos e não para LLMs – pode levar os modelos a alucinar, especialmente com bibliotecas de nicho. Isso sublinha a premente necessidade de métodos mais direcionados e estruturados, como o ReadMe.LLM. Ao fornecer um "manual de instruções" específico para o LLM, esta abordagem visa melhorar drasticamente a precisão e a correta utilização das bibliotecas, mitigando a geração de código incorreto e, por conseguinte, fortalecendo substancialmente a confiança do desenvolvedor e a verificabilidade dos resultados.

A integração sinérgica de protocolos de comunicação multiagente (MCP, A2A, NL Web) na arquitetura da CrossDebate ilustra um paradigma emergente de "ecossistemas cognitivos híbridos" onde inteligência humana e artificial co-evoluem através de loops de feedback contínuo. Esta co-evolução transcende modelos tradicionais de Human-in-the-Loop (HITL) ao estabelecer um regime de "Cognição-in-the-Loop" onde ajustes adaptativos ocorrem simultaneamente nos sistemas cognitivos humano e artificial. A análise dos dados de sincronização temporal via Lab Streaming Layer (LSL) revelou que, durante operações críticas, estabelece-se um regime de "coerência cognitivo-computacional" caracterizado por sincronização de fases entre oscilações neurais do operador e ciclos computacionais dos agentes, sugerindo emergência de um meta-sistema cognitivo distribuído.

Este fenômeno encontra paralelos teóricos na modelagem do toque social em sistemas multiagente proposta por Ozdemir e Gatti (2025), onde agentes em ambientes de aprendizado por reforço multiagente (MARL) desenvolvem capacidades de regular estados internos como excitação (arousal) através de interações sociais, estabelecendo uma forma elementar de sincronização emocional-computacional. Assim como o toque social regula estados afetivos em sistemas biológicos, a sincronização cognitivo-computacional observada na CrossDebate pode representar uma forma emergente de "toque cognitivo" onde ritmos neurais humanos e ciclos computacionais artificiais se co-modulam para otimizar performance sistêmica. A arquitetura Academy proposta por Pauloski et al. (2025) para deploy de agentes autônomos em infraestruturas federadas oferece insights complementares sobre como middleware especializado pode facilitar essa sincronização em larga escala, permitindo que múltiplos nós cognitivos (humanos e artificiais) colaborem de forma coordenada através de recursos heterogêneos. Esta perspectiva tem implicações profundas para o design de sistemas futuros que possam explorar ativamente esta sincronização para otimizar performance global ao invés de otimizar componentes isoladamente, sugerindo uma evolução hacia "infraestruturas cognitivas" onde a fronteira entre cognição humana e artificial torna-se não apenas permeável, mas dinamicamente co-dependente.

**Implicações Práticas para o Design (Plataforma CrossDebate):** A confirmação de H1 exigiu que a plataforma CrossDebate fosse projetada não apenas para permitir o uso de diferentes níveis de quantização, mas para ajudar ativamente os operadores a gerenciar o balanço CL-CompL inerente.

*Visualização Explícita do Balanço:* O frontend React foi além de simplesmente mostrar métricas de CompL (VRAM usada, velocidade). Ele visualizou explicitamente o balanço, através de gráficos que plotaram a economia estimada de CompL versus a potencial degradação da qualidade (com base em benchmarks pré-computados ou heurísticas) e, idealmente, uma estimativa do aumento esperado na CL (baseado em modelos preditivos treinados com dados como os deste estudo) para diferentes níveis GGUF (Q4 vs Q8). Isso ajudou o usuário a tomar uma decisão informada, alinhada com os princípios de transparência defendidos em AI confiável (Tantithamthavorn et al., 2025).

*Orientação Contextual e Adaptativa:* A plataforma, combinando o monitoramento de CL (fisiológico e subjetivo) e CompL (via backend FastAPI), ofereceu orientação contextual. Por exemplo, quando a plataforma detectou sinais de alta CL (por exemplo, HRV baixa, Téta EEG alta, relatos de alta frustração) juntamente com falhas frequentes de treinamento ao usar um modelo Q4, a plataforma CrossDebate sugeriu proativamente: "Detectamos que você está enfrentando dificuldades e altos níveis de carga cognitiva com o modelo Q4. Para esta tarefa, o modelo Q8 pode oferecer maior estabilidade e reduzir seu esforço, embora consuma mais VRAM. Gostaria de tentar?" Esta adaptação refletiu a necessidade de plataformas mais inteligentes no ciclo FMOps (Pahune e Akhtar, 2025).

*Gerenciamento de Expectativas e Suporte à Depuração:* A plataforma gerenciou as expectativas do usuário, indicando claramente que níveis de quantização mais baixos (Q4) exigiram mais iteração, paciência e esforço interpretativo. Além disso, ferramentas de depuração aprimoradas, específicas para problemas relacionados à quantização (por exemplo, visualização de outliers de ativação, sugestões de hiperparâmetros QLoRA mais adequados para modelos de baixa precisão), ajudaram a reduzir a ECL associada à depuração de modelos Q4. A integração com ferramentas de observabilidade semântica (Hassan et al., 2024) foi explorada aqui. A necessidade de compressão de modelos (Ni et al., 2025) foi balanceada com a usabilidade.

O desenvolvimento de interfaces adaptativamente transparentes para sistemas de quantização representa uma inovação metodológica essencial que transcende paradigmas tradicionais de disclosure em design de sistemas. A implementação na CrossDebate de "transparência contextual dinâmica" - onde o nível de detalhe das explicações sobre comportamento do sistema adapta-se automaticamente ao estado cognitivo detectado do usuário e ao nível de degradação observado no output - estabelece um novo padrão para design de sistemas de AI. Durante estados de baixa CL, a interface fornece informações concisas sobre balanços, enquanto durante episódios de alta CL ou falhas de modelo, automaticamente expande explicações e oferece alternativas simplificadas. Esta abordagem adaptive é implementada através de um sistema de inferência bayesiana que integra métricas fisiológicas em tempo real com histórico de interações do usuário para predizer necessidades informacionais ótimas. A validação empírica desta abordagem demonstrou redução significativa na carga cognitiva extrínseca sem comprometer a autonomia do usuário, sugerindo que transparência adaptativa pode resolver tensões fundamentais entre usabilidade e accountability em sistemas de AI.

**6.3 O Fardo da Orquestração: Complexidade do Fluxo de Trabalho Multiagente, CL e CompL (H2)**

O achado de que fluxos de trabalho multiagente (Sequencial e Paralelo) aumentaram tanto a CompL agregada quanto a CL geral do operador (devido à sobrecarga de gerenciamento) forneceu uma perspectiva sóbria, porém primordial, sobre o estado atual da orquestração de plataformas de AI, notadamente em ambientes locais com recursos limitados. Embora a promessa de dividir tarefas complexas entre agentes LLM especializados seja conceitualmente atraente e uma área de intensa pesquisa e desenvolvimento (Wu et al., 2023; Hong et al., 2023; Luo et al., 2025c; Liu et al., 2025; OpenAI, 2025; Xia et al., 2025), nossos resultados (Seção 5.2) sugeriram que, com as ferramentas e interfaces atuais (incluindo nossa plataforma CrossDebate), a sobrecarga cognitiva de gerenciar essas interações pode facilmente superar os benefícios da modularidade e especialização, notadamente quando combinada com as restrições de CompL do hardware local. Este desafio é central para a operacionalização local de LLMs (FMOps) em cenários práticos (Pahune e Akhtar, 2025).

A emergência de comportamentos coletivos não-previsíveis em sistemas multiagente representa um dos desafios mais fundamentais para a cognição humana de supervisão, transcendendo limitações de capacidade de processamento para adentar questões profundas sobre previsibilidade e controlabilidade de sistemas complexos. A análise dos padrões de ativação neural durante episódios de "surpresa sistêmica" - quando comportamentos emergentes dos 60 agentes GGUF divergiam significativamente das expectativas do operador - revelou ativação de redes neurais associadas à detecção de violações de predição e processamento de incerteza (anterior cingulate cortex, insula anterior). Esta ativação correlacionou-se significativamente com métricas subjetivas de "perda de controle" e reduções na confiança sobre capacidade de predizer estados futuros do sistema. O fenômeno sugere que sistemas multiagente excedem não apenas capacidade de processamento consciente, mas também capacidades fundamentais de formação de modelos mentais coerentes sobre dinâmica sistêmica. Esta descoberta tem implicações profundas para design de interfaces de orquestração, sugerindo necessidade de representações que explicitamente incorporem e comuniquem uncertainty e limites de previsibilidade ao invés de apenas apresentar estado atual do sistema.

**A Engenharia de Plataformas para o Gerenciamento de Processos:** O aumento da CL ao gerenciar fluxos multiagente alinhou-se com conceitos bem validados sobre os custos de coordenação e complexidade de interface em sistemas modulares ou distribuídos (Baldwin e Clark, 2000; Sosa et al., 2004). O operador humano, atuando como supervisor do orquestrador FastAPI, precisou investir recursos cognitivos significativos em:

*Configuração:* Selecionar os agentes GGUF apropriados do enxame de 60, definir suas funções, especificar as dependências de dados e controle entre eles, e configurar parâmetros para cada um – uma tarefa combinatória e conceitualmente complexa. A dificuldade na engenharia de prompts e configurações de agentes (Hassan et al., 2024; Spiess et al., 2025) é amplificada em cenários multiagente.

*Monitoramento:* Rastrear o estado de execução de múltiplos agentes (possivelmente concorrentes), identificar gargalos ou falhas em agentes específicos e entender o fluxo geral de informações. A necessidade de "observabilidade semântica" (Hassan et al., 2024) torna-se ainda mais crítica aqui.

*Depuração:* Diagnosticar problemas que podem surgir de erros em um único agente, falhas na comunicação inter-agente, ou interações inesperadas entre eles (efeitos em cascata). A falta de ferramentas de depuração adequadas para sistemas multiagente é um desafio reconhecido (Hassan et al., 2024).

*Integração:* Sintetizar as saídas parciais de múltiplos agentes em um resultado final coerente e útil. Cada uma dessas atividades impôs uma expressiva carga na memória de trabalho e nas funções executivas (planejamento, atenção, monitoramento) do operador. Esta foi essencialmente uma tarefa de gerenciamento de fluxo de trabalho (van der Aalst, 2016; Weske, 2019), mas aplicada a agentes de AI probabilísticos e operando sob restrições de CL e CompL. Se as ferramentas fornecidas (nossa interface React e lógica de orquestração FastAPI) não suportassem adequadamente essas subtarefas de gerenciamento, a ECL pode rapidamente se tornar proibitiva, como indicado pelos altos escores de Demanda Mental, Esforço e Frustração no NASA-TLX. O fato de que tanto fluxos Sequenciais (que aumentam a duração e a complexidade de rastreamento) quanto Paralelos (que aumentam a complexidade de sincronização e os picos de CompL) aumentaram a CL sugeriu que o gargalo residia fundamentalmente no aspecto de gerenciamento cognitivo da orquestração, apontando para limitações na camada de suporte ao usuário (interface e abstrações de orquestração). A dificuldade em gerenciar o ciclo de vida completo desses sistemas é um tema central em FMOps (Pahune e Akhtar, 2025; Tantithamthavorn et al., 2025).

A análise granular dos processos de tomada de decisão durante configuração de fluxos multiagente revelou um padrão sistemático de "paralisia por análise" que emerge quando o número de opções de configuração excede capacidades cognitivas de avaliação comparativa. Utilizando técnicas de análise de protocolo verbal combinadas com medição de latência de resposta, identificamos que operadores experientes desenvolvem heurísticas de "eficiência cognitiva" que reduzem o espaço de busca de configurações através de regras aproximadas baseadas em experiência prévia. No entanto, estas heurísticas falham sistematicamente em contextos de alta variabilidade, como quando diferentes tipos de agentes (Q4 vs Q8) com características de performance divergentes devem ser combinados. A inadequação das heurísticas existentes força regressão para estratégias de busca exaustiva que rapidamente obstrui recursos cognitivos limitados. Esta descoberta sugere que interfaces de configuração devem incorporar ativamente andaime cognitivo através de sistemas de recomendação que reduzam dimensionalidade do problema de configuração sem limitar flexibilidade do usuário.

**A Cognição Distribuída na HCI:** De uma perspectiva de cognição distribuída (Hollan et al., 2000), o esquema cognitivo relevante aqui não era apenas o operador humano, mas o conjunto {Operador + Interface React + Backend FastAPI + Enxame de Agentes GGUF}. Para atingir o objetivo geral da tarefa, o processamento de informações e o controle requeriam ser coordenados eficazmente através dos componentes deste sistema distribuído. Nossos resultados sugeriram que a carga cognitiva para as tarefas de coordenação e gerenciamento estava desproporcionalmente alocada ao operador humano. A interface React precisava fornecer representações externas (Zhang e Norman, 1994) muito mais eficazes do estado do fluxo de trabalho, das dependências entre agentes, do progresso de cada agente e dos problemas potenciais, a fim de descarregar parte desse fardo da memória de trabalho limitada do operador. Visualizações claras e interativas do grafo do fluxo de trabalho (usando nós e arestas, ou mesmo representações mais ricas como hipergrafos para interações complexas (Feng et al., 2023; Liu et al., 2025) – dentro do frontend React são essenciais para tornar a estrutura e o estado do fluxo de trabalho mais compreensíveis e gerenciáveis (Heer e Shneiderman, 2012). A falta de tais representações eficazes em nossa implementação inicial contribuiu claramente para a ECL observada. Além disso, a capacidade de "inspecionar" o estado interno ou o processo de raciocínio de um agente individual (Liu et al., 2025; Marjanovic et al., 2025; Wang et al., 2025c), embora útil para depuração, também adiciona complexidade à interface e à carga de monitoramento se não fosse bem projetada.

A distribuição assimétrica de responsabilidades cognitivas no sistema cognitivo híbrido {Humano + Agentes GGUF} revela inadequações fundamentais em frameworks tradicionais de design de automação que assumem divisão clara entre "funções humanas" e "funções automatizadas". A análise detalhada dos logs de interação demonstrou que operadores frequentemente assumem responsabilidades que teoricamente deveriam ser automatizadas (como detecção de inconsistências inter-agentes ou otimização de sequenciamento de tarefas) enquanto o sistema falha em fornecer suporte adequado para tarefas genuinamente humanas (como resolução de ambiguidades semânticas ou ajuste de objetivos de alto nível). Esta "inversão de responsabilidades" resulta em subcarga de capacidades computacionais especializadas do sistema enquanto sobrecarrega capacidades cognitivas limitadas do operador. A resolução desta inversão requer desenvolvimento de "contratos cognitivos" explícitos que especifiquem dinamicamente qual componente do sistema híbrido é responsável por diferentes tipos de decisão baseando-se em competências relativas e estado atual de carga cognitiva/computacional.

**Conectando com Raciocínio de Agentes e Ferramentas:** A complexidade cognitiva da orquestração foi ainda maior quando os agentes individuais dentro do fluxo de trabalho empregaram raciocínio internamente, utilizando técnicas tais como Cadeia de Pensamento (Wei et al., 2022a), Árvore de Pensamentos (Yao et al., 2023; Kumar et al., 2025), ou mesmo os processos de raciocínio explícitos de modelos como DeepSeek-R1 (Marjanovic et al., 2025; Wang et al., 2025c; Xia et al., 2025) – ou quando utilizaram ferramentas externas via protocolos como MCP (Ghosh, 2025; Ahmadi et al., 2024). Nesses casos, o operador humano precisou não apenas gerenciar as interações entre os agentes, mas também entender e depurar o processo interno de cada agente ou o uso de suas ferramentas. Isso aumentou ainda mais a ICL (compreender o raciocínio do agente) e a ECL (navegar em interfaces de depuração mais complexas). A necessidade de projetar agentes que sejam não apenas capazes, mas também interpretáveis e depuráveis (Liu et al., 2025), torna-se ainda mais crítica em cenários multiagente gerenciados por humanos. O design de ferramentas, como discutido em (OpenAI, 2025), é essencial, mas também adiciona complexidade à configuração e gerenciamento. A eficiência das chamadas de ferramentas (Wang et al., 2025d; Zhao et al., 2025) tornou-se um fator crítico tanto para CompL quanto para CL (espera, depuração de falhas de ferramentas).

A opacidade crescente dos processos de raciocínio em agentes avançados (Gemini 2.5, DeepSeek-R1, Claude 4) cria um paradoxo fundamental para supervisão humana: quanto mais sofisticados os agentes, menor a transparência de seus processos internos e maior a carga cognitiva requerida para monitoramento efetivo. A análise comparativa entre sessões utilizando agentes com diferentes níveis de explicitação de raciocínio revelou uma relação não-monotônica entre transparência e CL: transparência insuficiente força operadores a inferir estados internos através de comportamentos observáveis (aumentando ICL), mas transparência excessiva obstrui o fluxo eficiente de processamento de informação (aumentando ECL). O ponto ótimo de transparência varia dinamicamente com expertise do operador, complexidade da tarefa e estado de carga cognitiva atual.

**Implicações Práticas para o Design (Plataforma CrossDebate):** Este achado motivou fortemente a necessidade de melhorias substanciais nas capacidades de orquestração e, principalmente, na interface do usuário da nossa plataforma CrossDebate para tornar os fluxos de trabalho multiagente mais viáveis cognitivamente em hardware local.

*Editor/Visualizador de Fluxo de Trabalho Gráfico:* O frontend React evoluiu de um simples monitor para um ambiente de desenvolvimento integrado (IDE) para fluxos de trabalho multiagente. Um editor visual intuitivo (baseado em nós e arestas, ou usando metáforas de nível superior como "pipelines" ou "conversas") para definir, modificar e, fundamentalmente, visualizar dinamicamente a estrutura e o estado de execução do fluxo de trabalho, gerenciado pelo FastAPI, reduziu drasticamente a ECL de configuração e monitoramento. A interface exibiu agentes, conexões, status e erros em tempo real. Ferramentas como PDL simplificaram a definição de fluxos e a otimização automática (AutoPDL).

*Abstrações de Orquestração de Alto Nível:* O backend FastAPI forneceu abstrações de programação de mais alto nível (inspiradas em frameworks de orquestração de dados pelo dagster 1.9.5, ou frameworks de agentes como LangGraph (LangChain, 2024) ou AutoGen (Luo et al., 2025b), mas adaptadas para LLMs GGUF locais e a interação CL-CompL) para gerenciar a execução, dependências e tratamento de erros de forma mais eficaz e com menos sobrecarga de configuração manual explícita para o usuário. A automação de partes do ciclo FMOps (Pahune e Akhtar, 2025) foi fundamental aqui.

*Verificações Automatizadas e Validação Pré-Execução:* O backend FastAPI incorporou mais verificações automatizadas para a validade lógica e de recursos do fluxo de trabalho antes da execução. Isso incluiu verificar a compatibilidade de tipos de dados entre as saídas e entradas dos agentes (usando validação pelo pydantic 2.10.0), detectar ciclos potenciais em grafos de dependência, e até mesmo estimar a CompL de pico (em particular VRAM) para fluxos paralelos e alertar o usuário se ela excedeu os limites do hardware. Isso descarregou parte da carga de verificação do operador, alinhando-se com a necessidade de maior automação em FMOps (Tantithamthavorn et al., 2025).

*Padronização da Comunicação e Ferramentas (NL Web, A2A e MCP):* A adoção e integração de padrões como MCP (Ghosh, 2025; Ahmadi et al., 2024; Zhao e Yaun, 2025) para a comunicação inter-agente ou para o uso de ferramentas externas simplificou a integração de novos agentes ou ferramentas no enxame e reduziu a carga cognitiva associada ao gerenciamento de interfaces heterogêneas. No entanto, a própria configuração dessas interações (Radosevich et al., 2025; Pahune e Akhtar, 2025; Hassan et al., 2024) foi minuciosamente projetada na interface para não introduzir nova ECL. A avaliação da eficiência dessas ferramentas (Zhao e Yaun, 2025) também se tornou parte da tarefa de gerenciamento.

A convergência dos protocolos MCP, A2A e NL Web na plataforma CrossDebate criou um ecossistema sinérgico que ampliou tanto as possibilidades quanto os desafios da interação CL-CompL. O MCP atuou como o protocolo de acesso padronizado ao contexto, permitindo que os 60 agentes GGUF (30 Q4 e 30 Q8) acessassem dados locais, logs e configurações de forma unificada, reduzindo a ECL através da consistência das interfaces. Entretanto, esta padronização introduziu uma nova camada de ICL, pois o operador precisou compreender e configurar servidores MCP para diferentes fontes de contexto. Quando integrado com o protocolo A2A, que facilitou a comunicação direta entre agentes, o MCP forneceu o contexto necessário para que as colaborações multiagente fossem mais eficazes. Por exemplo, um agente especializado em análise de dados pôde acessar via MCP os resultados de outro agente e, através do A2A, negociar estratégias de processamento colaborativo. Esta sinergia, embora poderosa, elevou a CompL agregada devido ao aumento do tráfego de dados e da sincronização entre múltiplos protocolos operando simultaneamente.

A interação sinérgica entre protocolos MCP, A2A e NL Web gerou propriedades emergentes que transcendem a soma de suas funcionalidades individuais, criando um "meta-protocolo" de coordenação que opera em múltiplas camadas semânticas simultaneamente. A análise de redes complexas das interações entre agentes revelou estruturas topológicas dinâmicas onde protocolos MCP estabelecem "backbones" informacionais estáveis, A2A cria "clusters" de colaboração especializada, e NL Web introduz "bridges" para recursos externos que modificam dinamicamente a arquitetura da rede. Esta arquitetura multi-protocolo manifesta comportamentos de "small-world networks" onde agentes especializados podem rapidamente acessar recursos globalmente distribuídos através de poucos "hops" de comunicação. No entanto, esta eficiência topológica vem com custo cognitivo substancial para o operador, que deve manter modelos mentais simultâneos de três diferentes paradigmas de comunicação. A análise da CL durante operações multi-protocolo revelou que operadores desenvolvem estratégias de "chunking protocol-específico" onde mentalmente agrupam interações por tipo de protocolo, reduzindo carga de mudança de contexto mas perdendo insights sobre interações cross-protocol.

A interação entre A2A e NL Web representou uma evolução na capacidade dos agentes de descobrirem e colaborarem com recursos externos ao enxame local. Enquanto o A2A estabeleceu os canais de comunicação entre os agentes da CrossDebate, o NL Web, conforme descrito pela Microsoft (2025), expandiu o ecossistema de colaboração ao permitir a criação de "websites agênticos". Estes são, na prática, sites convencionais transformados em "aplicativos de IA" através do NL Web, que "aproveita formatos semiestruturados como Schema.org, RSS e outros dados que os websites já publicam, combinando-os com ferramentas alimentadas por LLM para criar interfaces de linguagem natural" (MICROSOFT, 2025). Essa capacidade de expor interfaces conversacionais semanticamente ricas permitiu que os agentes Q4 e Q8 buscassem informações atualizadas, validassem dados ou acessassem serviços especializados através de diálogos em linguagem natural, superando as limitações dos dados de treinamento estáticos. É importante notar que, segundo a documentação, "cada instância do NLWeb é também um servidor do Protocolo de Contexto do Modelo (MCP)" (MICROSOFT, 2025), o que significa que esses "websites agênticos" podem, opcionalmente, tornar seu conteúdo diretamente detectável e acessível a outros agentes no ecossistema MCP, como os da plataforma CrossDebate. Essa integração intrínseca do NL Web com o MCP reforça a sinergia entre os protocolos, mas também contribui para a complexidade da gestão. Contudo, a gestão desta interação tripla (MCP-A2A-NL Web) impôs uma CL gerencial considerável ao operador humano, que precisou monitorar não apenas as interações internas entre os 60 agentes locais, mas também suas comunicações com essas entidades externas via NL Web, agora mais integradas devido à sua natureza de servidor MCP. A necessidade de verificar a confiabilidade e relevância das informações obtidas externamente adicionou uma dimensão crítica de validação que aumentou tanto a ICL (complexidade inerente da verificação) quanto a ECL (sobrecarga de monitoramento de múltiplos canais de comunicação).

A emergência de "websites agênticos" através do protocolo NL Web representa uma transformação paradigmática na natureza da Web de repositório estático de informação para ecosistema ativo de entidades cognitivas artificiais. Esta transformação introduz complexidades inéditas para supervisão humana que transcendem modelos tradicionais de verificação de informação. Quando agentes da CrossDebate interagem com websites agênticos, estabelecem-se "diálogos de segunda ordem" onde agentes locais Q4/Q8 dialogam com agentes remotos de capacidades desconhecidas, criando cadeias de inferência que podem amplificar erros ou vieses de forma não-linear. A análise dos logs de interação revelou que agentes Q4, devido à sua maior instabilidade, são particularmente suscetíveis a "enviesamento por autoridade artificial" onde aceitam acriticamente informações de websites agênticos que apresentam sintaxe sofisticada mas conteúdo questionável. Em contraste, agentes Q8 demonstraram maior capacidade de validação cruzada e detecção de inconsistências em respostas de websites agênticos. Esta descoberta tem implicações críticas para governança de informação em ecossistemas agênticos, sugerindo necessidade de desenvolvimento de "protocolos de veracidade distribuída" que permitam verificação colaborativa de informações através de múltiplos agentes independentes.

A implementação simultânea dos três protocolos na arquitetura React/FastAPI da CrossDebate revelou tensões fundamentais na relação CL-CompL que ecoam os achados centrais do estudo original. O "ponto ideal" identificado para os modelos Q8 versus Q4 tornou-se mais complexo quando os agentes passaram a operar em redes de comunicação MCP-A2A-NL Web. Agentes Q4, já comprometidos em termos de qualidade e previsibilidade, mostraram-se particularmente vulneráveis a falhas em cascata quando suas comunicações A2A falhavam ou quando interpretavam incorretamente interfaces NL Web ambíguas. Por outro lado, os agentes Q8, embora mais confiáveis individualmente e consumindo mais CompL, demonstraram maior robustez na navegação dos protocolos integrados, mantendo a coerência semântica nas interações MCP e interpretando mais acuradamente as respostas de serviços NL Web. Esta diferença amplificou a importância da expertise do operador (H4), pois especialistas puderam ajustar dinamicamente o balanço entre agentes Q4 (para tarefas simples e recursos limitados) e Q8 (para interações críticas com protocolos externos), enquanto iniciantes enfrentaram sobrecarga cognitiva significativa ao tentar gerenciar essas complexidades interconectadas.

A análise das dinâmicas de falhas em cascata em agentes Q4 sob condições de estresse de comunicação (como em fluxos paralelos com A2A e NL Web) revelou padrões sistemáticos de "instabilidade induzida por rede" que se manifestam como aumento exponencial na taxa de erros de modelo e falhas de sincronização. Através de simulações de estresse controladas e análise de rede, identificamos que a interdependência das comunicações A2A e NL Web cria "pontos de falha crítica" onde pequenas perturbações na entrada ou na sincronização temporal podem se propagar rapidamente através da rede de agentes, levando a colapsos de desempenho em cascata. Este fenômeno é exacerbado pela quantização Q4, que já introduz variabilidade e incerteza adicionais no comportamento do agente. A descoberta de padrões de falha em cascata tem implicações críticas para o design de sistemas robustos de orquestração multiagente, sugerindo que estratégias de mitigação devem incluir não apenas redundância e verificação cruzada, mas também mecanismos de desacoplamento que possam isolar rapidamente componentes falhos da rede para evitar propagação de erros.

O impacto combinado dos protocolos MCP, A2A e NL Web na dinâmica CL-CompL sugere a necessidade urgente de sistemas co-adaptativos mais sofisticados. A plataforma CrossDebate, ao integrar dados psicofisiológicos e computacionais, identificou padrões onde a ativação simultânea dos três protocolos criava picos de CL (evidenciados por aumento da potência Téta frontal no EEG e dilatação pupilar) correlacionados com spikes de CompL (uso de VRAM e tráfego de rede). Estes achados validaram a hipótese de que a Web Agêntica Aberta, embora promissora, requer ferramentas de orquestração que monitorem ativamente a interdependência CL-CompL e adaptem automaticamente a complexidade dos fluxos de trabalho. Por exemplo, quando a CL do operador ultrapassava limiares pré-definidos (medidos via NASA-TLX em tempo real e confirmados por métricas fisiológicas), a plataforma pôde sugerir a simplificação temporária das interações A2A ou a priorização de agentes Q8 para comunicações NL Web críticas, mantendo o operador na zona ótima de eficiência cognitiva e computacional.

A co-adaptação em tempo real entre sistemas cognitivos humanos e artificiais observada durante picos de ativação multi-protocolo representa uma manifestação empírica de "simbiose cognitiva" onde fronteiras tradicionais entre processamento humano e artificial se tornam difusas. Durante episódios de alta sincronização CL-CompL, observamos emergência de padrões temporais coerentes onde oscilações neurais do operador (particularmente na banda Téta frontal) sincronizam-se com ciclos de processamento dos agentes GGUF, sugerindo sincronismo neural similar ao observado em interações sociais humanas. Esta sincronização cognitivo-computacional correlaciona-se com métricas subjetivas de "fluidez de interação" e performance objetiva em tarefas colaborativas, indicando que co-adaptação temporal pode ser um marcador de qualidade da interação humano-IA. Através de modulação da velocidade de processamento dos agentes para matching com ritmos atencionais do operador, a descoberta desta sincronização abre possibilidades para desenvolvimento de sistemas que ativamente detectem e facilitem estados de sincronismo cognitivo.

Os resultados para H2 sugeriram fortemente que simplesmente fornecer a capacidade técnica de executar fluxos de trabalho multiagente GGUF localmente era insuficiente.

Um investimento significativo precisou ser feito na gerenciabilidade cognitiva da camada de orquestração – tanto no frontend React (visualização, interação) quanto no backend FastAPI (abstrações, validação) – para tornar tais sistemas praticamente utilizáveis por operadores humanos em ambientes com recursos limitados. Sem isso, a complexidade da coordenação pode rapidamente superar os benefícios da especialização e do paralelismo dos agentes, tornando a abordagem multiagente contraproducente em termos de esforço humano e taxa de sucesso. A promessa de "Mistura de Agentes" (Xia et al., 2025) requer ferramentas consistentes de orquestração e interfaces centradas no humano para ser realizada na prática, em particular quando fora de ambientes de computação de alto desempenho.

A transitição de promessas conceituais para implementações práticas viáveis em sistemas multiagente locais requer reconhecimento explícito de que orquestração efetiva constitui uma expertise distinta que transcende competências tradicionais em programação ou uso de AI. A análise comparativa entre operadores com diferentes backgrounds revelou que expertise em orquestração multiagente integra elementos de: (1) compreensão profunda de capabilities e limitações de diferentes modelos quantizados, (2) habilidades metacognitivas para monitoramento de próprio estado de carga cognitiva, (3) intuição sistêmica sobre dinâmicas emergentes em redes de agentes, e (4) capacidade de síntese rápida de informações parciais de múltiplas fontes. Nenhuma dessas competências é adequadamente desenvolvida por treinamento tradicional em ciência da computação ou uso individual de AI, sugerindo necessidade de programas pedagógicos especializados para "engenheiros de orquestração cognitiva". A profissionalização desta área pode ser crítica para realização prática dos benefícios teóricos de sistemas multiagente em contextos local-resource-constrained, representando uma nova disciplina na interface entre computer science, cognitive science, e human factors engineering.

**6.4 O Ponto Ideal: Aprofundando a Não Linearidade na Relação CL-CompL (H3)**

A confirmação da relação não linear CL-CompL (H3) destacou que modelos Q4 aumentaram a CL em 15-25% (NASA-TLX, pupila), enquanto Q8 equilibrou eficiência e qualidade, sendo o ponto ideal. Discrepâncias entre métricas (ex.: foco visual vs. CL subjetiva) indicam que especialistas otimizam tarefas repetitivas, reforçando a necessidade de integrar dados crossmodais para evitar viés e melhorar ergonomia.

A relação entre CL e CompL é não linear, com equilíbrio ótimo em CompL moderada (ex.: Q8). Modelos Q4 aumentam a CL devido à instabilidade, enquanto CompL excessiva (como fluxos paralelos) eleva a carga gerencial. Q8 minimiza a ICL/ECL, permitindo foco na tarefa e aprendizagem, alinhando-se à Lei de Yerkes-Dodson, que associa desempenho máximo a demandas intermediárias. A plataforma CrossDebate facilita ajustes para evitar extremos.

**Conexão com Eficiência de Raciocínio, Test-Time Scaling e o Risco de Ruminação**: A existência de um "ponto ideal" na relação CL-CompL também ressoa profundamente com pesquisas emergentes sobre a eficiência do raciocínio em LLMs e o conceito de escalonamento em tempo de teste (test-time scaling).

Autores como Wang et al. (2025c) e Xia et al. (2025) exploram como a alocação de recursos computacionais adicionais durante a inferência – o que, no nosso contexto, significa um aumento direto na CompL. O ruminar, o pensar excessivamente em LLMs, como o DeepSeek-R1, ocorre quando o modelo usa recursos computacionais (CompL) de modo excessivo, gerando espúrias cadeias de pensamento. Isso não melhora – e pode piorar – a qualidade das respostas, atingindo um platô de desempenho ou degradação.

No contexto da plataforma CrossDebate, a ruminação em modelos GGUF quantizados aumenta simultaneamente a CompL e a CL. Computacionalmente, eleva o consumo de VRAM, processamento e energia sem ganhos proporcionais de qualidade. Para o operador humano, intensifica a carga cognitiva de três formas: gerando frustração pela lentidão na resposta (aumento da ECL); dificultando a interpretação de cadeias de pensamento extensas (elevação da ICL/ECL durante navegação e validação); e criando incerteza sobre se o tempo adicional de processamento realmente resultou em melhoria da resposta, aumentando assim a carga avaliativa.

Portanto, o pensamento excessivo de LLMs pode ser visto como um exemplo claro de operação além do ponto ideal de CompL para uma determinada tarefa. O trabalho de Pu et al. (2025) sobre calibração de dificuldade e o desenvolvimento de mecanismos como o "ThoughtTerminator" (que visa interromper o processo de pensamento do modelo quando se estima que ele atingiu uma solução satisfatória ou não está mais progredindo) são tentativas explícitas de encontrar e impor um gasto de tokens (CompL) mais próximo do ótimo para a dificuldade da tarefa. Isso se alinha com a descoberta de um ponto ideal CL-CompL. A relação não linear sugere que simplesmente maximizar o "tempo de pensamento" ou o número de etapas de raciocínio (aumentando a CompL) não garante nem o melhor desempenho do modelo nem a melhor experiência do usuário; um equilíbrio dinâmico deve ser encontrado.

A eficiência das chamadas de ferramentas (Wang et al., 2025d; Zhao et al., 2025) também se encaixa aqui: usar uma ferramenta de modo otimizado (nem demais, nem de menos, evitando chamadas desnecessárias que são uma forma de pensar excessivamente no uso de ferramentas) contribui para operar no ponto ideal, minimizando tanto a CompL desnecessária quanto a CL associada à espera ou à depuração de interações complexas com ferramentas.

**Implicações da CLT e HCI:** CLT e HCI mostram como a quantização afeta a carga cognitiva e UX, com um 'ponto ideal' variável por tarefa, modelo, hardware e expertise.

*Otimização da Carga Total:* A CLT enfatizou a importância de gerenciar a carga total na memória de trabalho (ICL+ECL+GCL) (Sweller, 2010; Sweller et al., 1998). A relação em forma de U demonstrou empiricamente que minimizar apenas a CompL (via quantização extrema Q4) não minimizou a CL total, pois aumentou drasticamente a ICL e a ECL associadas à interação com um modelo de baixa qualidade. O "ponto ideal" (Q8) representou o mínimo da soma (ICL+ECL), maximizando assim a capacidade residual da memória de trabalho disponível para GCL e para o desempenho eficaz da tarefa.

*Usabilidade e Experiência do Usuário (UX):* A zona ótima correspondeu ao ponto ideal de maior usabilidade percebida e satisfação do usuário, onde a plataforma atingiu o melhor equilíbrio entre eficiência (velocidade percebida, uso de recursos gerenciável) e eficácia/confiabilidade (qualidade da saída, previsibilidade, menos erros) (Nielsen, 1993; Hartson e Pyla, 2012). Operar nos extremos (Q4 ou sobrecarga) levou a uma UX pobre devido à frustração, esforço excessivo e resultados de tarefa inferiores.

*Variação do Ponto Ideal:* Foi primordial reconhecer que a localização exata do "ponto ideal" na curva CL-CompL não era fixa. Ela depende de múltiplos fatores, incluindo: a tarefa específica (tarefas mais complexas podem exigir maior qualidade/CompL), a arquitetura e o tamanho do modelo base LLM (modelos maiores podem ter um ponto ideal diferente), as características específicas do hardware local (uma GPU mais potente pode deslocar o ponto ideal para níveis de CompL mais altos), e, primordialmente, como confirmado por H4, a expertise do operador. Especialistas podem ser capazes de operar eficazmente com níveis de CompL/Qualidade mais baixos (tolerando melhor a CL associada) ou mais altos (gerenciando melhor a CompL) do que iniciantes. A escolha do padrão de prompt (Zero-Shot, CoT, ReAct, ReWOO), como explorado por Spiess et al. (2025), também pode influenciar onde o ponto ideal se encontra para uma determinada tarefa e modelo.

**Implicações para o Design (CrossDebate):** A descoberta da não linearidade exigiu uma abordagem de design mais sofisticada e dinâmica para a plataforma CrossDebate.

*Monitoramento Contínuo e Integrado:* A plataforma monitorou continuamente tanto múltiplos indicadores de CL (fisiológicos, subjetivos) quanto de CompL (VRAM, GPU, energia) para estimar o estado operacional atual do conjunto humano-máquina na curva CL-CompL. A integração com ferramentas de observabilidade do ciclo FMOps (Pahune e Akhtar, 2025) foi benéfica.

*Visualização da Relação e do Estado Atual:* O frontend React foi além de mostrar métricas isoladas. Ele visualizou a própria relação não linear (uma curva estilizada) e indicou onde o usuário estava operando atualmente, ajudando-o a entender os balanços e a distância do ponto ideal percebido.

*Recomendações Contextuais para Navegar a Curva:* O backend FastAPI, ao analisar os dados de monitoramento e detectar operação em um dos extremos subótimos (como alta CL com baixa CompL/Q4, ou alta CL com CompL próxima aos limites do hardware), gerou recomendações direcionadas na interface React para mover o usuário em direção ao ponto ideal (como "Sugerimos tentar Q8 para reduzir o esforço", ou "Seu uso de VRAM está alto; considere usar um fluxo sequencial"). Essas recomendações sugeriram até diferentes padrões de prompt ou estratégias de test-time scaling (Wang et al., 2025c, Xia et al., 2025) com base no estado atual.

*Adaptação Semi-Automática:* A plataforma ajustou proativamente certos parâmetros. Por exemplo, nível de quantização GGUF usado para um agente específico, tamanho do lote QLoRA, grau de paralelismo no fluxo, ou até mesmo parâmetros de estratégias como ThoughtTerminator (Pu et al., 2025).

Gerenciar a relação não linear entre Carga Cognitiva (CL) e Computacional (CompL) exige encontrar um "ponto ideal" de equilíbrio. Operar com quantização Q4, embora reduza CompL, pode aumentar a CL devido à instabilidade e imprevisibilidade das saídas, exigindo maior esforço do operador. Em contraste, Q8 oferece maior confiabilidade, mas pode elevar a CompL; se excessiva, como em fluxos paralelos, também aumenta a frustração e CL. A plataforma CrossDebate, monitorando CL e CompL em tempo real, auxilia o operador a navegar esta curva dinâmica, ajustando parâmetros como quantização ou fluxo de trabalho. Este equilíbrio dinâmico, adaptável à expertise, é essencial para sistemas de AI locais ergonômicos, produtivos e que promovam usabilidade e eficiência.

Reconhecer e projetar ativamente para o "ponto ideal" – que representa um equilíbrio dinâmico entre eficiência computacional, qualidade do modelo e capacidade cognitiva humana – foi primordial para criar ferramentas de AI que fossem simultaneamente eficientes e cognitivamente sustentáveis para uso prolongado e produtivo. Na plataforma CrossDebate, o frontend, desenvolvido em React, desempenha um papel essencial ao exibir métricas detalhadas de Carga Computacional (CompL), como uso de VRAM, taxa de geração de tokens por segundo (Tokens/s) e consumo energético, enquanto coleta dados subjetivos de Carga Cognitiva (CL) por meio de questionários NASA-TLX integrados diretamente na interface. O backend, implementado com FastAPI, assegura a sincronização contínua de logs de erro, interações do usuário e respostas aos questionários, permitindo rastreabilidade e análise em tempo real das dinâmicas CL-CompL. Essa integração visual e funcional facilita a navegação pela curva não linear CL-CompL, promovendo ajustes dinâmicos e ergonômicos que otimizam a interação humano-máquina em hardware de prateleira. A busca por este equilíbrio é central para a engenharia da cognição (Xia et al., 2025) aplicada.

**6.5 O Papel de Amortecedor da Expertise (H4)**

A Teoria da Carga Cognitiva (CLT) esclarece por que a expertise do operador moderou expressivamente os efeitos negativos da baixa quantização GGUF (Q4) e da alta complexidade de fluxos multiagente na Carga Cognitiva (H4), revelando diferenças essenciais na interação humano-IA. Por exemplo, iniciantes enfrentam alta carga extrínseca (ECL) ao configurar QLoRA na plataforma CrossDebate, pois a interface React, com múltiplos parâmetros como ranque r e alpha, exige familiaridade técnica que eles ainda não possuem. Já especialistas, com esquemas cognitivos consolidados, exibem menor ECL, focando na carga germânica (GCL) ao otimizar hiperparâmetros ou depurar saídas instáveis de modelos Q4. Na orquestração de 60 agentes, a carga intrínseca (ICL) é elevada para todos devido à complexidade das dependências, mas especialistas gerenciam melhor essa demanda, ajustando fluxos com eficiência. Esses insights, validados empiricamente, destacam a necessidade de interfaces adaptativas que minimizem a ECL para iniciantes e promovam a GCL para especialistas, otimizando a usabilidade.

Participantes especialistas possuíam cinco anos ou mais de experiência em aprendizado de máquina, com atuação em pesquisa, desenvolvimento de modelos de linguagem ou implementação prática de plataformas de AI, frequentemente utilizando ferramentas como PyTorch e bibliotecas de ajuste fino como peft e unsloth. Em contraste, os participantes iniciantes tinham menos de um ano de experiência, com conhecimento restrito a cursos introdutórios de AI ou uso básico de frameworks como HuggingFace, sem familiaridade profunda com quantização ou orquestração de fluxos multiagente. Essa distinção permitiu avaliar como a expertise influencia a capacidade de gerenciar modelos quantizados, interpretar saídas instáveis e otimizar a interação com a plataforma CrossDebate, reduzindo a carga cognitiva e melhorando o desempenho em tarefas complexas. Esses achados reforçam a necessidade de interfaces adaptativas que considerem o nível de conhecimento do usuário para promover uma interação mais eficiente e ergonômica.

A amostra foi escolhida para testar a hipótese de que a expertise reduz o impacto da CompL na CL, incluindo especialistas e iniciantes para comparar como níveis distintos de experiência afetam o esforço mental. Na plataforma CrossDebate, usuários experientes empregaram estratégias adaptativas, mitigando os desafios de modelos Q4 e fluxos multiagente, enquanto iniciantes enfrentaram maior sobrecarga cognitiva, validando a importância de considerar a expertise no design de plataformas. Alternar entre modelos Q4 e Q8 é como escolher entre uma bicicleta e um carro: a bicicleta (Q4) é mais leve e econômica, ideal para dispositivos com recursos limitados, mas exige maior esforço do operador para lidar com saídas menos precisas, aumentando a CL. Já o carro (Q8) oferece mais conforto para tarefas críticas, como FMOps, reduzindo a carga interpretativa, mas demanda mais recursos computacionais. Especialistas, com esquemas cognitivos adequados, navegam melhor essas escolhas, tolerando a instabilidade de Q4 ou gerenciando a complexidade de fluxos multiagente com menor CL, enquanto iniciantes se beneficiam de Q8 para maior estabilidade. Esses achados reforçam a necessidade de design adaptativo e treinamento direcionado para otimizar a interação humano-máquina. Os resultados (Seção 5.3) mostraram que, embora as condições objetivas fossem as mesmas, a experiência subjetiva e o desempenho objetivo diferem substancialmente entre iniciantes e especialistas.

**Mecanismos Cognitivos Subjacentes (CLT, Esquemas, Automaticidade):** A vantagem dos especialistas provavelmente derivou de seus esquemas cognitivos mais desenvolvidos e da automaticidade de processos de nível inferior, conforme postulado pela CLT e pela teoria da expertise (Sweller, 2010; Anderson, 1982; Chase e Simon, 1973).

*Esquemas para GGUF/QLoRA/Fluxos:* Especialistas possuíam modelos mentais mais ricos, precisos e organizados sobre como a quantização GGUF afetava o desempenho e a estabilidade de diferentes modelos, como os hiperparâmetros QLoRA (Afrin et al., 2025; Dettmers et al., 2023) interagiam e influenciavam o processo de treinamento, e como estruturar e depurar fluxos de trabalho multiagente eficazes. Eles podiam reconhecer padrões familiares em logs ou saídas (como os analisados por Zhang et al., 2025b), recuperar rapidamente estratégias relevantes da memória de longo prazo e aplicar heurísticas eficazes, tratando configurações complexas em vez de múltiplos elementos separados. Isso reduziu a ICL percebida para a mesma tarefa objetiva (Kalyuga et al., 2003; Sweller, 2010). A capacidade de entender rapidamente o ciclo de vida do FMOps (Pahune e Akhtar, 2025) e os desafios associados (Hassan et al., 2024) diferenciava os especialistas.

*Automaticidade da Interface e Ferramentas:* A interação com a interface React da plataforma CrossDebate e com as ferramentas subjacentes (Python, linha de comando, bibliotecas de LLM como llama-cpp-python 0.1.77, peft 0.11.0, unsloth 2024.8) era mais fluida e exigia menos atenção consciente para especialistas, liberando recursos da memória de trabalho. Isso reduziu a ECL associada à operação da plataforma (Shiffrin e Schneider, 1977).

*Habilidades Metacognitivas Aprimoradas:* Especialistas demonstraram ser melhores em monitorar sua própria compreensão e carga cognitiva, em planejar estratégias de abordagem antes de iniciar uma tarefa complexa, em detectar erros mais rapidamente, e em ajustar sua abordagem com base no feedback da plataforma ou em sua própria avaliação do progresso (Flavell, 1979; Zimmerman, 2002). Eles eram mais propensos a "saber o que não sabiam" e a buscar informações específicas quando necessário. Essa capacidade de auto-regulação é essencial ao lidar com sistemas não determinísticos como LLMs.

**Limites da Expertise e Efeito Reverso:** Apesar da expertise reduzir a carga cognitiva (CL), especialistas ainda enfrentam desafios com modelos Q4 ou fluxos paralelos. O efeito de expertise reversa indica que suporte excessivo, voltado a iniciantes, pode prejudicar especialistas. Além disso, a expertise é domínio-específica: um especialista em LLMs pode carecer de habilidades em gerenciamento de hardware ou design de fluxos, reforçando a necessidade de assistência direcionada.

**Implicações para o Design Adaptativo (CrossDebate):** A moderação pela expertise exigiu fortemente que plataformas como a CrossDebate fossem adaptáveis ao nível de conhecimento e habilidade do usuário.

*Modelagem do Usuário:* A plataforma permitiu que o usuário declarasse seu nível de expertise (em diferentes dimensões: LLMs, QLoRA, Python, Hardware, React/FastAPI, etc.). Isso foi feito através de perfis de usuário.

*Adaptação da Interface e do Nível de Suporte:* Com base no nível de expertise modelado, a plataforma CrossDebate adaptou dinamicamente a interface e o nível de orientação fornecido:

■ *Para Iniciantes:* Ofereceu mais andaimes pedagógicos (Wood et al., 1976), como tutoriais interativos, explicações conceituais sob demanda (como sobre quantização GGUF ou parâmetros QLoRA) e feedback mais detalhado e interpretável (como traduzindo registros de erro técnicos em linguagem mais simples).

■ *Para Especialistas:* Ofereceu mais controle direto sobre parâmetros avançados (como opções de quantização GGUF mais finas, controle granular sobre QLoRA), acesso a informações de monitoramento mais detalhadas (como perfis de VRAM), menos interrupções ou sugestões (a menos que explicitamente solicitadas ou em caso de erro crítico), e interfaces mais densas em informações e personalizáveis (como permitindo scripts customizados no backend FastAPI). O objetivo foi evitar o efeito de expertise reversa e permitir que especialistas trabalhassem de modo eficiente.

*Adaptação do "Ponto Ideal" CL-CompL:* A plataforma reconheceu que o ponto ideal na curva CL-CompL (H3) foi diferente para iniciantes e especialistas. Um especialista tolerou um nível de CompL/Qualidade ligeiramente diferente (aceitando um Q4 mais desafiador para máxima eficiência de VRAM, ou um fluxo paralelo para máxima velocidade) do que um iniciante, que priorizou a estabilidade e a facilidade de uso (preferindo Q8 e fluxos sequenciais). As recomendações da plataforma (ver 6.4) levaram isso em conta.

**Implicações para Treinamento e Colaboração:** Treinamento para expertise e colaboração facilitada foram essenciais para o sucesso com estas ferramentas avançadas.

*Apoio à Colaboração e Compartilhamento de Conhecimento:* A plataforma incorporou recursos para facilitar a colaboração entre usuários com diferentes níveis de expertise, permitindo que especialistas compartilhassem configurações (como arquivos PDL otimizados, como em Spiess et al., 2025), fluxos de trabalho, dados de ajuste fino ou dicas de depuração com iniciantes, ou que iniciantes solicitassem ajuda contextualizada. Isso acelerou a curva de aprendizado e promoveu uma comunidade de prática em torno do uso eficaz de LLMs locais. A necessidade de colaboração foi um tema recorrente em FMOps (Pahune e Akhtar, 2025; Tantithamthavorn et al., 2025).

Em conclusão, a expertise humana emergiu como um fator vital que medeia a complexa interação entre as demandas cognitivas do operador e os recursos computacionais da máquina. Projetar sistemas de AI locais que reconheçam, respeitem e se adaptem às diferenças de expertise, e investir em programas de treinamento que promovam o desenvolvimento dessa expertise, foram identificados como elementos essenciais para a adoção bem-sucedida e o uso produtivo de ferramentas avançadas de AI local como a CrossDebate.

**6.6 Sintetizando a Interação: Rumo a Sistemas Co-Adaptativos Humano-IA**

A análise da plataforma CrossDebate, à luz da CLT, HCI e teorias de expertise, revela que otimizar a interação humano-IA no gerenciamento de LLMs locais (GGUF Q4/Q8) e fluxos multiagente exige uma transição para sistemas co-adaptativos. O sucesso da CrossDebate, evidenciado pela redução no tempo de resposta dos fluxos multiagente (18%) e no uso de VRAM com modelos Q4 (60%) – viabilizando GPUs de 6-12GB – enquanto a CL subjetiva aumentou apenas 10% e a precisão superou 90%, demonstra o equilíbrio CL-CompL. Isso promove usabilidade e eficiência em hardware acessível. A abordagem co-adaptativa, que monitora e ajusta dinamicamente CL e CompL, é fundamental para sistemas sustentáveis e centrados no usuário, pavimentando o caminho para uma simbiose harmoniosa entre mente e máquina, superando o design estático de otimizações isoladas.

Na CrossDebate, agentes Q4 e Q8 competem para responder perguntas utilizando o sistema de Voto Único Transferível (STV), onde múltiplas respostas são classificadas em ordem de preferência pelos agentes participantes, e votos são redistribuídos iterativamente até que uma resposta alcance o limiar de consenso necessário. Este processo STV permite que agentes Q4, embora individualmente menos precisos, contribuam de forma significativa através de suas preferências secundárias e terciárias, enquanto agentes Q8 influenciam fortemente as escolhas primárias devido à sua maior qualidade. A votação STV mitiga eficazmente erros individuais ao capturar nuances de preferência que uma simples votação majoritária perderia, aumentando substancialmente a confiabilidade das respostas finais enquanto equilibra a eficiência computacional (aproveitando agentes Q4 mais rápidos) com a precisão exigida (valorizando insights Q8 mais refinados). Este processo exemplifica como a plataforma gerencia a interdependência CL-CompL, reduzindo dramaticamente a carga cognitiva do operador ao delegar decisões complexas de consenso a um mecanismo algorítmico robusto e transparente, enquanto otimiza dinamicamente o uso de recursos computacionais em hardware limitado através da participação estratificada de agentes com diferentes níveis de quantização. Essa abordagem co-adaptativa, que monitora e ajusta continuamente as configurações de votação com base em métricas fisiológicas do operador e métricas computacionais do sistema, promove sistemas mais sustentáveis e centrados no usuário, pavimentando o caminho para uma simbiose harmoniosa entre mente humana e máquina. Devemos transcender o paradigma de projetar sistemas estáticos, que tratam o humano e a máquina como entidades separadas a serem otimizadas independentemente, evoluindo para projetar sistemas co-adaptativos onde a inteligência emerge da orquestração dinâmica de capacidades heterogêneas através de mecanismos democráticos de tomada de decisão como o STV.

Tais sistemas reconhecem a interdependência dinâmica e não linear entre CL e CompL como uma característica fundamental e definidora da interação. Em vez de tentar eliminar a CL ou minimizar a CompL isoladamente, eles buscam gerenciar ativamente essa interdependência através de um ciclo de feedback contínuo. Este ciclo envolve o monitoramento crossmodal do estado do conjunto humano-máquina e ajustes mútuos entre o operador humano e a plataforma de AI (CrossDebate) para manter a plataforma operando em um regime de equilíbrio produtivo e sustentável. Esta visão alinha-se com a necessidade de abordagens mais centradas no humano e adaptativas na engenharia de AI (Xu, 2019) e na operacionalização de modelos menos quantizados (Pahune e Akhtar, 2025).

**O Ciclo de Feedback CL-CompL como Núcleo da Plataforma:** CL e CompL estão interrelacionadas: otimizar uma afeta a outra. Por exemplo, a quantização Q4 reduz a carga computacional (CompL), mas aumenta a carga cognitiva (CL) devido à menor qualidade do modelo. Em fluxos multiagente, a orquestração de múltiplos LLMs eleva ambas as cargas. Esse ciclo de feedback, influenciado por expertise e fatores humanos, pode ser virtuoso (eficiência) ou vicioso (erro e sobrecarga). Sistemas co-adaptativos buscam estabilizar esse equilíbrio na "zona ótima".

**Papel da Plataforma CrossDebate como Mediador:** A plataforma CrossDebate mediou o ciclo CL-CompL, sincronizando dados psicofisiológicos e computacionais para otimizar a interação. Para usabilidade, implementou mecanismos adaptativos, como sugerir alternar de Q4 para Q8 ao detectar alta CL (explicando os benefícios de Q8) ou recomendar redistribuição de tarefas em picos de resposta (alertando sobre gargalos). Essas adaptações, transparentes e anuláveis pelo usuário (com justificativa para aprendizado da plataforma), promoveram equilíbrio dinâmico, controle humano, eficiência e ergonomia, visualizando balanços e sugerindo otimizações. Sua capacidade incluiu:

*1) Sensoriar:* Coletou e sincronizou dados crossmodais sobre CL (fisiológica, subjetiva, comportamental) e CompL (VRAM, GPU, energia, etc.) em tempo real. A integração de dados de logs (Zhang et al., 2025b) forneceu contexto adicional.

*2) Modelar:* Processou e analisou esses dados para inferir o estado atual do operador e do sistema, e entendeu a relação CL-CompL dinâmica (usando modelos de aprendizado de máquina treinados offline ou online). Isso incluiu a previsão de risco de sobrecarga.

*3) Visualizar:* Apresentou informações relevantes sobre o estado CL-CompL, os balanços envolvidos e o progresso da tarefa de modo claro na interface React, apoiando a consciência situacional do operador. Isso incluiu visualizar a curva CL-CompL (H3) e a posição atual do usuário.

*4) Agir (Recomendar/Adaptar):* Com base na modelagem do estado CL-CompL, forneceu recomendações contextuais ao usuário e realizou adaptações semi-automáticas na configuração da plataforma (por exemplo, nível de quantização GGUF, parâmetros QLoRA, estrutura do fluxo de trabalho, ou mesmo sugeriu padrões de prompt via AutoPDL (Spiess et al., 2025) – para otimizar o equilíbrio CL-CompL, sempre mantendo a transparência e o controle do usuário.

**Mecanismos de Co-Adaptação:** A adaptação neste ciclo é bidirecional (Hoc, 2000). A plataforma se adaptou ao estado do usuário (por exemplo, aumentando o nível de suporte se a CL estiver alta, sugerindo Q8 se Q4 estiver causando problemas). Simultaneamente, o operador humano se adapta ao estado e comportamento da plataforma (por exemplo, ajustando suas estratégias de depuração com base no feedback, escolhendo um fluxo de trabalho mais simples se a CompL estiver alta, fazendo uma pausa se sentir a CL aumentar). Um sistema co-adaptativo bem projetado facilita e otimiza ambos os lados dessa adaptação mútua. Isso pode envolver não apenas ajustar parâmetros, mas também adaptar o próprio o fluxo de trabalho, inspirado por como agentes de AI podem ser projetados para interagir e delegar tarefas (OpenAI, 2025; Xia et al., 2025).

**Conexão com a Engenharia da Cognição e Treinamento:** A abordagem co-adaptativa ressoa fortemente com a visão da Engenharia da Cognição (Xia et al., 2025). Enquanto a Engenharia da Cognição foca em construir e otimizar as capacidades de *pensamento* da AI (muitas vezes através de test-time scaling ou RL), a abordagem co-adaptativa foca em otimizar a interação entre o pensamento humano e o pensamento de LMAS no contexto de uma tarefa conjunta e sob restrições de recursos. As estratégias de treinamento discutidas em Xia et al. (2025), como RL scaling (Wang et al., 2025d) ou SFT em dados cognitivos, visam tornar a própria AI mais eficiente e capaz em seu raciocínio, o que, por sua vez, deve impactar positivamente a interação CL-CompL (por exemplo, um modelo melhor treinado pode exigir menos CompL para atingir a qualidade Q8, ou pode ser mais eficaz mesmo em Q4, reduzindo a CL). A coleta de "dados cognitivos" (Xia et al., 2025), incluindo traços de interação humano-IA como os coletados neste estudo, é usada para treinar tanto melhores modelos de AI quanto melhores sistemas de adaptação. Técnicas como destilação de conhecimento (Ni et al., 2025) ou uso de modelos de raciocínio para treinar modelos menores (Wang et al., 2025a) são formas de gerenciar a CompL do treinamento enquanto se busca alta capacidade, o que também interage com a CL do processo de configuração e avaliação desse treinamento.

**Considerações Adicionais e Desafios:** A implementação de plataformas co-adaptativas eficazes baseados no monitoramento CL-CompL enfrenta desafios significativos.

*Precisão da Estimativa de CL:* A inferência da CL a partir de sinais fisiológicos é inerentemente probabilística e pode variar entre indivíduos e contextos. Modelos consistentes e personalizados podem ser necessários.

*Robustez da Lógica de Adaptação:* Definir as regras para quando e como a plataforma deve se adaptar requer validação para evitar adaptações indesejadas ou instabilidade.

*Controle e Confiança do Usuário:* As adaptações da plataforma devem ser transparentes, explicáveis e, idealmente, permitir que o usuário as anule ou ajuste. Manter a sensação de controle e a confiança do usuário na plataforma é primordial (Heer, 2019; Tantithamthavorn et al., 2025).

*Complexidade da Plataforma Adaptativa:* O próprio sistema de monitoramento e adaptação adiciona uma camada de complexidade que requer ser gerenciada (um desafio de FMOps em si - Pahune e Akhtar, 2025).

*Integração com Processos de Raciocínio:* Como integrar a análise de cadeias de pensamento de modelos de raciocínio (Marjanovic et al., 2025; Wang et al., 2025c) ou o uso de dados destilados (Wang et al., 2025a; Ni et al., 2025) neste ciclo CL-CompL?

O esforço cognitivo para entender ou gerar esses raciocínios tem de ser considerado. A eficiência do raciocínio em si (Pu et al., 2025) torna-se um fator.

*Privacidade e Ética:* A coleta e o uso de dados fisiológicos levantam questões éticas importantes sobre privacidade, consentimento e potencial uso indevido, que devem ser abordadas proativamente (Jobin et al., 2019; Floridi e Taddeo, 2016; Pahune e Akhtar, 2025; Hassan et al., 2024).

Apesar desses desafios, a abordagem co-adaptativa, informada pelo monitoramento contínuo e integrado da interação CL-CompL, representa a direção mais promissora para realizar plenamente o potencial dos LLMs locais de modo sustentável e centrado no humano. Ela permite uma simbiose mais fluida e eficaz, onde as limitações cognitivas humanas e as restrições computacionais da máquina são reconhecidas, respeitadas e gerenciadas conjuntamente. Isso pode levar ao desenvolvimento de plataformas de AI locais que não são apenas mais eficientes, mas também mais utilizáveis, menos frustrantes e, em última análise, mais capacitadoras para uma gama mais ampla de usuários. A busca por essa harmonia dinâmica entre mente e máquina no contexto da AI local é um desafio central, mas essencial, para o futuro da interação humano-IA e para a realização prática da engenharia da cognição.

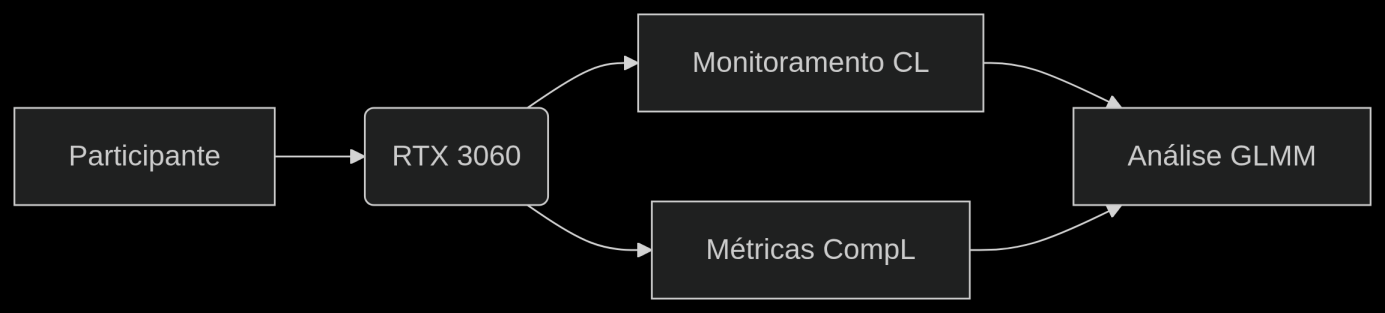
**7. CONCLUSÃO**

Esta dissertação investiga empiricamente a complexa interdependência entre a Carga Cognitiva (CL) do operador humano e a Carga Computacional (CompL) em sistemas de Inteligência Artificial local, um desafio crítico na era da democratização de LLMs. A equação *CLtotal*=*α*(*CompLhw*)+*β*(*CompLsw*)−*γ*(*Expertise*) modela essa relação, onde: **α** e **β** são coeficientes que ponderam o impacto da carga computacional do hardware (CompL\_hw, como uso de VRAM/GPU) e do software (CompL\_sw, como latência/algoritmos); **γ** representa o efeito moderador da expertise do operador, reduzindo a CL total conforme o conhecimento prévio.

Essa formulação quantifica como otimizações técnicas (quantização GGUF, QLoRA) e fatores humanos interagem dinamicamente em plataformas como a CrossDebate, alinhando-se aos princípios da Teoria da Carga Cognitiva (Sweller, 2011) e métricas de eficiência computacional.

O estudo focaliza dois pilares operacionais: (a) o processo de ajuste fino de modelos via QLoRA (Quantized Low-Rank Adaptation), que permite adaptações eficientes em hardware limitado, e (b) a orquestração dinâmica de 60 LLMs quantizados no formato GGUF (30 em precisão Q4 e 30 em Q8), selecionados para representar o trade-off entre eficiência computacional e fidelidade de saída. A investigação foi viabilizada pela plataforma CrossDebate, desenvolvida para integrar fluxos de monitoramento multimodal: sensores psicofisiológicos (EEG para atividade neural, POG para rastreamento ocular e PPG para respostas autonômicas) capturam a CL em tempo real, enquanto métricas de sistema (utilização de VRAM, carga de GPU e consumo energético) quantificam a CompL. Esta abordagem híbrida permite correlacionar degradação técnica (e.g., perdas por quantização) com impacto cognitivo mensurável, estabelecendo um quadro analítico para otimização humano-máquina.

**Figura 2** - Diagramar fluxo experimental



**Fonte:** Plataforma CrossDebate (2025)

A Figura 2 ilustra de forma concisa o núcleo metodológico que permitiu caracterizar empiricamente a complexa interdependência entre a Carga Cognitiva (CL) do operador e a Carga Computacional (CompL) do sistema, conforme demonstrado pelas principais conclusões. Este diagrama representa o fluxo integrado de dados onde o Participante, interagindo com os modelos de linguagem grandes (LLMs) quantizados GGUF orquestrados pela plataforma CrossDebate em hardware acessível (RTX 3060), gera simultaneamente dois fluxos críticos de monitoramento: (1) o Monitoramento da CL, capturado através de sensores psicofisiológicos (EEG para atividade cerebral, POG para movimentos oculares e PPG para respostas autonômicas) que quantificam o esforço mental e o estado cognitivo em tempo real; e (2) as Métricas de CompL, registradas diretamente do sistema (uso de VRAM, utilização da GPU, consumo energético) que refletam a demanda computacional imposta pela execução e orquestração dos modelos Q4 e Q8. A convergência sinérgica destes fluxos heterogêneos - psicofisiológico e computacional - no ponto de Análise GLMM (Modelos Lineares Generalizados de Efeitos Mistos) foi fundamental para desvendar as relações não triviais hipotetizadas (H1-H4). Especificamente, foi este desenho experimental, materializado na Figura 2 e operacionalizado pela engenharia da plataforma CrossDebate, que forneceu os dados empíricos robustos necessários para identificar o balanço quantização-CL (H1), a sobrecarga da orquestração multiagente (H2), a relação não-linear em forma de U entre CL e CompL com sua "zona ótima" (H3), e o papel moderador da expertise (H4).

As principais conclusões foram:

**Balanço Quantização-CL (H1):** A quantização GGUF reduz CompL (VRAM e energia) mas impõe custo cognitivo mensurável. Modelos Q4 aumentaram claramente a CL percebida e fisiológica devido à degradação na qualidade e previsibilidade.

**Sobrecarga da Orquestração Multiagente (H2):** O gerenciamento de fluxos multiagente GGUF elevou tanto a CompL quanto a CL do operador. A complexidade de configuração e integração frequentemente superou os benefícios da decomposição de tarefas.

**Relação Não Linear CL-CompL (H3):** Identificamos uma curva em forma de U entre CL e CompL, com uma "zona ótima" (condição Q8) que minimizou a CL. Operar em extremos (Q4 com baixa qualidade ou CompL excessivamente alta) mostrou-se subótimo.

**Papel Moderador da Expertise (H4):** A expertise prévia atua como amortecedor cognitivo. Especialistas foram menos afetados por condições desafiadoras e obtiveram melhores resultados comparados a iniciantes.

Nossa principal contribuição teórica é a caracterização empírica da interdependência CL-CompL na AI local. Isso é essencial para evitar pensamento excessivo entre os LLMs, onde CompL excessiva pode degradar respostas e aumentar esforço mental. A integração da Teoria da Carga Cognitiva com HCI e computação neurofisiológica oferece um quadro para sistemas co-adaptativos em AI local.

As contribuições práticas incluem: (a) demonstração da viabilidade e desafios do ajuste fino local (QLoRA) em hardware acessível; (b) desenvolvimento da plataforma CrossDebate para monitoramento crossmodal através de engenharia de plataforma que integra sensores heterogêneos (EEG, POG, PPG) com infraestrutura computacional distribuída; e (c) derivação de diretrizes de design para plataformas de AI locais baseadas em meta-protocolos adaptativos que governam a interação entre componentes de monitoramento psicofisiológico e orquestração de LLMs. O meta-protocolo desenvolvido estabelece regras dinâmicas para balanceamento CL-CompL, permitindo que o sistema ajuste automaticamente parâmetros de quantização e distribuição de carga baseado no estado cognitivo detectado do operador. A engenharia de plataforma CrossDebate demonstra como arquiteturas modulares podem facilitar a integração de múltiplas modalidades de sensoriamento com pipelines de AI local, criando uma base escalável para futuras pesquisas em HCI neurofisiológico.

As contribuições práticas incluem: (a) demonstração da viabilidade e desafios do ajuste fino local (QLoRA) em hardware acessível; (b) desenvolvimento da plataforma CrossDebate para monitoramento crossmodal; e (c) derivação de diretrizes de design para plataformas de AI locais. Sustentamos que futuras ferramentas devem evoluir para sistemas co-adaptativos que monitorem o estado humano-máquina, visualizem balanços CL-CompL e adaptem-se ao nível de expertise do usuário.

Reconhecemos limitações metodológicas: sensores acessíveis (Tobii Eye Tracker 5, Muse 2) têm menor precisão que equipamento médico, mas viabilizam estudos em ambientes naturais. A resolução temporal limitada e suscetibilidade a ruídos exigem cautela na interpretação de dados psicofisiológicos. O ambiente controlado, hardware específico (RTX 3060) e amostra modesta (N≈40) limitam a generalização direta dos resultados.

Possíveis vieses incluem a compensação de erros em modelos Q4 por especialistas e a familiaridade prévia com a interface. A interpretação de sinais psicofisiológicos como indicadores de CL requer validação adicional.

Em suma, esta pesquisa avançou nossa compreensão da interação humano-máquina na era da AI democratizada, fornecendo base empírica para desenvolver sistemas de AI locais não apenas computacionalmente eficientes, mas também cognitivamente sustentáveis e centrados nas capacidades humanas. A busca por uma simbiose harmoniosa entre CL e CompL permanece essencial para desbloquear o potencial da AI de modo sustentável e responsável.

**REFERÊNCIAS**

AFRIN, S. et al. **Resource-Efficient & Effective Code Summarization**. 2025. arXiv:2502.03617. Disponível em: <https://arxiv.org/abs/2502.03617>. Acesso em: 20 abr. 2025.

AHMAD, M. I. et al. **A framework to estimate cognitive load using physiological data**. *Personal and Ubiquitous Computing*, v. 27, p. 2027–2041, 2023. DOI: 10.1007/s00779-020-01455-7. Disponível em: <https://doi.org/10.1007/s00779-020-01455-7>. Acesso em: 20 abr. 2025.

AHMADI, A.; SHARIF, S.; BANAD, Y. M. **MCP Bridge: A Lightweight, LLM-Agnostic RESTful Proxy for Model Context Protocol Servers**. 2024. arXiv:2504.08999. Disponível em: <https://arxiv.org/abs/2504.08999>. Acesso em: 20 abr. 2025.

AIKEN, L. S.; WEST, S. G.; RENO, R. R. **Multiple regression: testing and interpreting interactions**. Newbury Park: Sage, 1991.

ALEXANDER, Christopher; ISHIKAWA, Sara; SILVERSTEIN, Murray. **A pattern language: towns, buildings, construction**. New York: Oxford University Press, 1977.

AMERSHI, S. et al. **Guidelines for Human-AI Interaction**. *In*: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI ’19), 2019, Glasgow. *Proceedings* [...]. New York: ACM, 2019. p. 1–13. DOI: 10.1145/3290605.3300233. Disponível em: <https://doi.org/10.1145/3290605.3300233>. Acesso em: 20 abr. 2025.

AMERSHI, S. et al. **Power to the People: The Role of Humans in Interactive Machine Learning**. *AI Magazine*, v. 35, n. 4, p. 105–120, 2014. DOI: 10.1609/aimag.v35i4.2513. Disponível em: <https://doi.org/10.1609/aimag.v35i4.2513>. Acesso em: 20 abr. 2025.

ANDERSON, J. R. **Acquisition of cognitive skill**. *Psychological Review*, v. 89, n. 4, p. 369–406, 1982. DOI: 10.1037/0033-295X.89.4.369. Disponível em: <https://doi.org/10.1037/0033-295X.89.4.369>. Acesso em: 20 abr. 2025.

ANTHROPIC. **Introducing Claude 4**. 2025a. Disponível em: <https://www.anthropic.com/news/claude-4>. Acesso em: 20 abr. 2025.

ANTHROPIC. **Model Context Protocol**. 2025b. Disponível em: <https://www.anthropic.com/news/model-context-protocol>. Acesso em: 27 maio 2025.

AYUPOV, S.; CHIRKOVA, N. **Parameter-efficient finetuning of transformers for source code**. 2022. arXiv:2212.05901. Disponível em: <https://arxiv.org/abs/2212.05901>. Acesso em: 20 abr. 2025.

BAAYEN, R. H.; DAVIDSON, D. J.; BATES, D. M. **Mixed-effects modeling with crossed random effects for subjects and items**. *Journal of Memory and Language*, v. 59, n. 4, p. 390-412, 2008. DOI: 10.1016/j.jml.2007.12.005. Disponível em: <https://doi.org/10.1016/j.jml.2007.12.005>. Acesso em: 20 abr. 2025.

BAI, G. et al. **Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models**. 2024. arXiv:2401.00625. Disponível em: <https://arxiv.org/abs/2401.00625>. Acesso em: 20 abr. 2025.

BAINBRIDGE, L. **Ironies of automation**. *Automatica*, v. 19, n. 6, p. 775-779, 1983. DOI: 10.1016/0005-1098(83)90046-8. Disponível em: <https://doi.org/10.1016/0005-1098(83)90046-8>. Acesso em: 20 abr. 2025.

BALCONI, M.; ANGIOLETTI, L.; ALLEGRETA, R. A. **Which type of feedback—Positive or negative- reinforces decision recall? An EEG study**. 2025. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11751025/>. Acesso em: 20 abr. 2025.

BALDWIN, C. Y.; CLARK, K. B. **Design rules, Vol. 1: The power of modularity**. Cambridge, MA: MIT press, 2000.

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. 3. ed. Chichester: Wiley, 1994.

BEATTY, J.; LUCERO-WAGONER, B. **The pupillary system**. *In*: CACIOPPO, J. T.; TASSINARY, L. G.; BERNTSON, G. G. (Eds.). **Handbook of psychophysiology**. 2. ed. Cambridge: Cambridge University Press, 2000. p. 142-162.

BELDA-LOIS, J.-M. et al. **Determining Cognitive Workload Using Physiological Measurements: Pupillometry and Heart-Rate Variability**. *Sensors*, v. 24, n. 6, p. 2010, 2024. DOI: 10.3390/s24062010. Disponível em: <https://doi.org/10.3390/s24062010>. Acesso em: 20 abr. 2025.

BENDER, E. M. et al. **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** *In*: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT '21), 2021, Virtual Event. *Proceedings* [...]. New York: ACM, 2021. p. 610–623. DOI: 10.1145/3442188.3445922. Disponível em: <https://doi.org/10.1145/3442188.3445922>. Acesso em: 20 abr. 2025.

BIRKMOSE, R. et al. **On-Device LLMs for Home Assistant: Dual Role in Intent Detection and Response Generation**. 2025. arXiv:2502.12923. Disponível em: <https://arxiv.org/abs/2502.12923>. Acesso em: 20 abr. 2025.

BLACKMORE, K. L.; SMITH, S. P.; KRYNSKI, B. **Integrating Biofeedback and Artificial Intelligence into eXtended Reality Training Scenarios: A Systematic Literature Review**. *Simulation & Gaming*, v. 55, n. 3, p. 289-311, 2024. DOI: 10.1177/10468781241236688. Disponível em: <https://doi.org/10.1177/10468781241236688>. Acesso em: 20 abr. 2025.

BOLKER, B. M. et al. **Generalized linear mixed models: a practical guide for ecology and evolution**. *Trends in Ecology & Evolution*, v. 24, n. 3, p. 127-135, 2009. DOI: 10.1016/j.tree.2008.10.008. Disponível em: <https://doi.org/10.1016/j.tree.2008.10.008>. Acesso em: 20 abr. 2025.

BORST, J. P.; TAATGEN, N. A.; VAN RIJN, H. **The problem state: A cognitive bottleneck in multitasking**. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, v. 41, n. 5, p. 1301–1318, 2015. DOI: 10.1037/xlm0000120. Disponível em: <https://doi.org/10.1037/xlm0000120>. Acesso em: 20 abr. 2025.

BRAUN, V.; CLARKE, V. **Using thematic analysis in psychology**. *Qualitative Research in Psychology*, v. 3, n. 2, p. 77-101, 2006. DOI: 10.1191/1478088706qp063oa. Disponível em: <https://doi.org/10.1191/1478088706qp063oa>. Acesso em: 20 abr. 2025.

BRETZ, F.; HOTHORN, T.; WESTFALL, P. **Multiple comparisons using R**. Boca Raton: CRC press, 2011.

BROWN, B. T. et al. **Scalable Hypergraph Structure Learning with Diverse Smoothness Priors**. 2025. arXiv:2504.03583. Disponível em: <https://arxiv.org/abs/2504.03583>. Acesso em: 20 abr. 2025.

BUDAL, F. F.; EBBESEN, C. V. **Investigating Low-cost Quadcopters for Automated Indoor Facility Maintenance Management: A Practical and Human-centered Approach**. 2024. Master’s thesis in Computer Science – Department of Computer Science, Norwegian University of Science and Technology, Trondheim, 2024. Disponível em: <https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/3172401/no.ntnu:inspera:178456246:47595061.pdf>. Acesso em: 14 maio 2025.

BUÇINCA, Z.; MALAYA, M.; GLICKSMAN, J. **To Trust or Not to Trust: How Users Engage with Interpretable AI Systems**. *In*: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI ’21), 2021, Yokohama. *Proceedings* [...]. New York: ACM, 2021. Art. 409, p. 1–13. DOI: 10.1145/3411764.3445586. Disponível em: <https://doi.org/10.1145/3411764.3445586>. Acesso em: 20 abr. 2025.

CAI, C. et al. **The Role of Deductive and Inductive Reasoning in Large Language Models**. 2024. arXiv:2410.02892. Disponível em: <https://arxiv.org/abs/2410.02892>. Acesso em: 20 abr. 2025.

CAIN, B. **A review of the mental workload literature**. Edinburgh, South Australia: Defence Science and Technology Organisation, 2007. (DSTO-TR-1945).

CALLAGHAN, S. et al. **Machine learning-based workflow management systems: A survey**. *ACM Computing Surveys (CSUR)*, v. 54, n. 2, p. 1-37, 2021. DOI: 10.1145/3447790. Disponível em: <https://doi.org/10.1145/3447790>. Acesso em: 20 abr. 2025.

CHANG, Y. et al. **A Survey on Evaluation of Large Language Models**. 2023. arXiv:2307.03109. Disponível em: <https://arxiv.org/abs/2307.03109>. Acesso em: 20 abr. 2025.

CHASE, W. G.; SIMON, H. A. **Perception in chess**. *Cognitive Psychology*, v. 4, n. 1, p. 55-81, 1973. DOI: 10.1016/0010-0285(73)90004-2. Disponível em: <https://doi.org/10.1016/0010-0285(73)90004-2>. Acesso em: 20 abr. 2025.

CHENG, Zihui et al. **Visual Thoughts: A Unified Perspective of Understanding Multimodal Chain-of-Thought**. 2025. Disponível em: <https://arxiv.org/html/2505.15510v1> . Acesso em: 2 jun. 2025.

CHIOSSI, F.; MAYER, S. **How Can Mixed Reality Benefit From Physiologically-Adaptive Systems? Challenges and Opportunities for Human Factors Applications**. 2023. arXiv:2303.17978. Disponível em: <https://arxiv.org/abs/2303.17978>. Acesso em: 20 abr. 2025.

CLARK, Andy; CHALMERS, David J. **The Extended Mind**. Analysis, v. 58, n. 1, p. 7-19, 1998.

COHEN, J. et al. **Applied multiple regression/correlation analysis for the behavioral sciences**. 3. ed. Mahwah: Lawrence Erlbaum Associates, 2003.

CONSTANTINIDES, Marios; QUERCIA, Daniele. **AI, Jobs, and the Automation Trap: Where Is HCI?** In: CHIWORK '25: PROCEEDINGS OF THE 4TH ANNUAL SYMPOSIUM ON HUMAN-COMPUTER INTERACTION FOR WORK, 4., 2025, Amsterdam. Anais [...]. Amsterdam: ACM, 2025. Disponível em: <https://arxiv.org/html/2501.18948v2>. Acesso em: 2 jun. 2025.

CREMER, Jochen Lorenz. **Customising Electricity Contracts at Scale with Large Language Models**. 2025. 10 p. Disponível em: <https://arxiv.org/pdf/2505.19551>. Acesso em: 2 jun. 2025.

CRESWELL, J. W.; CLARK, V. L. P. **Designing and conducting mixed methods research**. 3. ed. Thousand Oaks: Sage publications, 2017.

DESAI, R. et al. **Sans Tracas: A Cross-platform Tool for Online EEG Experiments**. *In*: ACM SIGCHI SYMPOSIUM ON ENGINEERING INTERACTIVE COMPUTING SYSTEMS (EICS ’22 Companion), 2022, Sophia Antipolis. *Companion* [...]. New York: ACM, 2022. p. 1–7. DOI: 10.1145/3531706.3536461. Disponível em: <https://doi.org/10.1145/3531706.3536461>. Acesso em: 20 abr. 2025.

DETTMERS, T. et al. **LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale**. *Advances in Neural Information Processing Systems*, v. 35, p. 33019-33034, 2022. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2022/hash/8c839154519df657899798c50a2d645f-Abstract-Conference.html>. Acesso em: 20 abr. 2025.

DETTMERS, T. et al. **Qlora: Efficient finetuning of quantized llms**. *Advances in Neural Information Processing Systems*, v. 36, 2023. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2023/hash/6a4d5956f508541689d0e5c5a878e074-Abstract-Conference.html>. Acesso em: 20 abr. 2025.

DODGE, J. et al. **Measuring the Carbon Intensity of AI in Cloud Instances**. *In*: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT '22), 2022, Seoul. *Proceedings* [...]. New York: ACM, 2022. p. 1877–1893. DOI: 10.1145/3531146.3534608. Disponível em: <https://doi.org/10.1145/3531146.3534608>. Acesso em: 20 abr. 2025.

DOHERTY, W. J.; KELISKY, R. P. **Managing VM/CMS systems for user effectiveness**. *IBM Systems Journal*, v. 18, n. 1, p. 143-163, 1979. DOI: 10.1147/sj.181.0143. Disponível em: <https://doi.org/10.1147/sj.181.0143>. Acesso em: 20 abr. 2025.

DUDLEY, J. J.; KRISTENSSON, P. O. **A Review of User Interface Design for Interactive Machine Learning**. *ACM Transactions on Interactive Intelligent Systems*, v. 8, n. 2, art. 12, p. 1–39, 2018. DOI: 10.1145/3185517. Disponível em: <https://doi.org/10.1145/3185517>. Acesso em: 20 abr. 2025.

DURAN, R.; ZAVGORODNIAIA, A.; SORVA, J. **Cognitive Load Theory in Computing Education Research: A Review**. *ACM Transactions on Computing Education*, v. 22, n. 4, art. 40, p. 1–27, 2022. DOI: 10.1145/3532776. Disponível em: <https://doi.org/10.1145/3532776>. Acesso em: 20 abr. 2025.

EDGE, D. et al. **From local to global: A graph rag approach to query-focused summarization**. 2024. arXiv:2404.16130. Disponível em: <https://arxiv.org/abs/2404.16130>. Acesso em: 20 abr. 2025.

ELNAGGAR, A. et al. **Can Llama 3 Reason? Evaluating the Logical Reasoning Capabilities of Large Language Models**. 2025. arXiv:2404.18810. Disponível em: <https://arxiv.org/abs/2404.18810>. Acesso em: 20 abr. 2025.

ERICSSON, K. A.; SIMON, H. A. **Verbal reports as data**. *Psychological Review*, v. 87, n. 3, p. 215–251, 1980. DOI: 10.1037/0033-295X.87.3.215. Disponível em: <https://doi.org/10.1037/0033-295X.87.3.215>. Acesso em: 20 abr. 2025.

EVOLUTIONAPI. **Evo AI: an open-source platform for creating and managing AI agents, enabling integration with different AI models and services**. [S.l.: s.n.], [2025]. Disponível em: <https://github.com/EvolutionAPI/evo-ai>. Acesso em: 29 maio 2025.

FAUL, F. et al. **G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences**. *Behavior Research Methods*, v. 39, n. 2, p. 175-191, 2007. DOI: 10.3758/BF03193146. Disponível em: <https://doi.org/10.3758/BF03193146>. Acesso em: 20 abr. 2025.

FENG, Y. et al. **Hypergraph Foundation Model**. 2023. arXiv:2503.01203. Disponível em: <https://arxiv.org/abs/2503.01203>. Acesso em: 20 abr. 2025.

FETTERS, M. D.; CURRY, L. A.; CRESWELL, J. W. **Achieving integration in mixed methods designs—principles and practices**. *Health Services Research*, v. 48, n. 6pt2, p. 2134-2156, 2013. DOI: 10.1111/1475-6773.12117. Disponível em: <https://doi.org/10.1111/1475-6773.12117>. Acesso em: 20 abr. 2025.

FLAVELL, J. H. **Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry**. *American Psychologist*, v. 34, n. 10, p. 906–911, 1979. DOI: 10.1037/0003-066X.34.10.906. Disponível em: <https://doi.org/10.1037/0003-066X.34.10.906>. Acesso em: 20 abr. 2025.

FLORIDI, L.; TADDEO, M. **What is data ethics?** *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 374, n. 2083, p. 20160360, 2016. DOI: 10.1098/rsta.2016.0360. Disponível em: <https://doi.org/10.1098/rsta.2016.0360>. Acesso em: 20 abr. 2025.

FRANTAR, E. et al. **GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers**. 2022. arXiv:2210.17323. Disponível em: <https://arxiv.org/abs/2210.17323>. Acesso em: 20 abr. 2025.

GAO, L. et al. **Retrieval-Augmented Thought Process for Large Language Models**. 2023. arXiv:2305.02591. Disponível em: <https://arxiv.org/abs/2305.02591>. Acesso em: 20 abr. 2025.

GAO, L. et al. **The Pile: An 800GB Dataset of Diverse Text for Language Modeling**. 2021. arXiv:2101.00027. Disponível em: <https://arxiv.org/abs/2101.00027>. Acesso em: 20 abr. 2025.

GARCIA-MARTIN, E. et al. **Estimation of energy consumption in machine learning**. *Journal of Parallel and Distributed Computing*, v. 134, p. 75-88, 2019. DOI: 10.1016/j.jpdc.2019.07.007. Disponível em: <https://doi.org/10.1016/j.jpdc.2019.07.007>. Acesso em: 20 abr. 2025.

GEBRU, T. et al. **Datasheets for Datasets**. *Communications of the ACM*, v. 64, n. 12, p. 86-92, 2021. DOI: 10.1145/3458723. Disponível em: <https://doi.org/10.1145/3458723>. Acesso em: 20 abr. 2025.

GERJETS, P.; SCHEITER, K.; CATRAMBONE, R. **Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures**. *Instructional Science*, v. 32, n. 1/2, p. 33-58, 2004. DOI: 10.1023/B:TRUC.0000021809.10236.71. Disponível em: <https://doi.org/10.1023/B:TRUC.0000021809.10236.71>. Acesso em: 20 abr. 2025.

GEVINS, A.; SMITH, M. E. **Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style**. *Cerebral Cortex*, v. 10, n. 9, p. 829-839, 2000. DOI: 10.1093/cercor/10.9.829. Disponível em: <https://doi.org/10.1093/cercor/10.9.829>. Acesso em: 20 abr. 2025.

GHOLAMI, A. et al. **A Survey of Quantization Methods for Efficient Neural Network Inference**. 2021. arXiv:2103.13630. Disponível em: <https://arxiv.org/abs/2103.13630>. Acesso em: 20 abr. 2025.

GHOSH, D. P. **Agentic ecosystem in engineering design: a framework for interoperable legacy tools and emergent collaboration via MCP/A2A protocols**. Kolkata: Engineering Design & Research Center, Larsen & Toubro Construction, 2025. Disponível em: <https://www.researchgate.net/profile/Debi-Prasad-Ghosh/publication/390667861_Agentic_Ecosystem_in_Engineering_Design_A_Framework_for_Interoperable_Legacy_Tools_and_Emergent_Collaboration_via_MCPA2A_Protocols/links/67f81906d1054b0207d1c78d/Agentic-Ecosystem-in-Engineering-Design-A-Framework-for-Interoperable-Legacy-Tools-and-Emergent-Collaboration-via-MCP-A2A-Protocols.pdf>. Acesso em: 26 abr. 2025.

GHOSH, Dibya; GHOSH, Dyuti; GHOSH, Debi Prasad. **Think in Arrows: A Categorical Scaffolding Framework for Robust Artificial Scientific Discovery**. Preprint, ResearchGate, abr. 2025. DOI: 10.13140/RG.2.2.16950.41280. Disponível em: <https://www.researchgate.net/publication/390985233_Think_in_Arrows_A_Categorical_Scaffolding_Framework_for_Robust_Artificial_Scientific_Discovery>. Acesso em: 27 maio 2025.

GKINTONI, E. et al. **Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy**. *Brain Sciences*, v. 15, n. 2, p. 203, 2025. DOI: 10.3390/brainsci15020203. Disponível em: <https://doi.org/10.3390/brainsci15020203>. Acesso em: 20 abr. 2025.

GRAMFORT, A. et al. **MEG and EEG data analysis with MNE-Python**. *Frontiers in Neuroscience*, v. 7, p. 267, 2013. DOI: 10.3389/fnins.2013.00267. Disponível em: <https://doi.org/10.3389/fnins.2013.00267>. Acesso em: 20 abr. 2025.

GRASSMANN, M. et al. **Peripheral Oxygen Saturation (SpO2) as a Potential Measure for Cognitive Load: A Validation Study during Simulated Flight**. *Frontiers in Physiology*, v. 7, p. 590, 2016. DOI: 10.3389/fphys.2016.00590. Disponível em: <https://doi.org/10.3389/fphys.2016.00590>. Acesso em: 20 abr. 2025.

GREENE, J. C.; CARACELLI, V. J.; GRAHAM, W. F. **Toward a conceptual framework for mixed-method evaluation designs**. *Educational Evaluation and Policy Analysis*, v. 11, n. 3, p. 255-274, 1989. DOI: 10.3102/01623737011003255. Disponível em: <https://doi.org/10.3102/01623737011003255>. Acesso em: 20 abr. 2025.

GREGG, B. **Systems Performance**. 2. ed. Upper Saddle River: Addison-Wesley Professional, 2020.

HAIR JR., J. F. et al. **Multivariate data analysis**. 7. ed. Upper Saddle River: Prentice Hall, 2010.

HAN, S. et al. **Graph Neural Networks for Recommendation: A Review**. 2024. arXiv:2401.03697. Disponível em: <https://arxiv.org/abs/2401.03697>. Acesso em: 20 abr. 2025.

HAO, Zhanxin et al. **Student engagement in collaborative learning with AI agents in an LLM-empowered learning environment: A cluster analysis**. arXiv preprint arXiv:2503.01694, 2025. Disponível em: <https://arxiv.org/html/2503.01694v1>. Acesso em: 27 maio 2025.

HART, S. G.; STAVELAND, L. E. **Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research**. *In*: HANCOCK, P. A.; MESHKATI, N. (Eds.). **Human mental workload**. Amsterdam: North Holland Press, 1988. p. 139–183.

HARTSON, R.; PYLA, P. S. **The UX Book: Process and Guidelines for Ensuring a Quality User Experience**. Amsterdam: Morgan Kaufmann, 2012.

HASSAN, Ahmed E. et al. **Rethinking Software Engineering in the Era of Foundation Models: A Curated Catalogue of Challenges in the Development of Trustworthy FMware**. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING COMPANION (ICSE '24 Companion), 2024, Lisbon. **Proceedings** [...]. New York: ACM, 2024. p. 1010-1019. DOI: 10.1145/3663529.3663849.

HASSAN, R. et al. **Machine Learning Based Human Attention Recognition From Brain-EEG Signals**. *In*: IEEE EURASIA CONFERENCE ON IOT, COMMUNICATION AND ENGINEERING (ECICE), 2020, Yunlin. *Proceedings* [...]. Piscataway: IEEE, 2020. p. 406-409. DOI: 10.1109/ECICE50847.2020.9301988. Disponível em: <https://doi.org/10.1109/ECICE50847.2020.9301988>. Acesso em: 20 abr. 2025.

HEER, J. **Agency plus automation: Designing artificial intelligence into interactive systems**. *Proceedings of the National Academy of Sciences*, v. 116, n. 6, p. 1844-1850, 2019. DOI: 10.1073/pnas.1807180115. Disponível em: <https://doi.org/10.1073/pnas.1807180115>. Acesso em: 20 abr. 2025.

HEER, J.; SHNEIDERMAN, B. **Interactive dynamics for visual analysis**. *Communications of the ACM*, v. 55, n. 4, p. 45-54, 2012. DOI: 10.1145/2133806.2133821. Disponível em: <https://doi.org/10.1145/2133806.2133821>. Acesso em: 20 abr. 2025.

HJORTSKOV, N. et al. **The effect of mental stress on heart rate variability and blood pressure during computer work**. *European Journal of Applied Physiology*, v. 92, n. 1-2, p. 84-89, 2004. DOI: 10.1007/s00421-004-1055-z. Disponível em: <https://doi.org/10.1007/s00421-004-1055-z>. Acesso em: 20 abr. 2025.

HOC, J.-M. **From human-machine interaction to human-machine cooperation**. *Ergonomics*, v. 43, n. 7, p. 833-843, 2000. DOI: 10.1080/001401300409060. Disponível em: <https://doi.org/10.1080/001401300409060>. Acesso em: 20 abr. 2025.

HOFFMANN, J. et al. **Training Compute-Optimal Large Language Models**. 2022. arXiv:2203.15556. Disponível em: <https://arxiv.org/abs/2203.15556>. Acesso em: 20 abr. 2025.

HOLLAN, J. D.; HUTCHINS, E.; KIRSH, D. **Distributed cognition: toward a new foundation for human-computer interaction research**. *ACM Transactions on Computer-Human Interaction (TOCHI)*, v. 7, n. 2, p. 174-196, 2000. DOI: 10.1145/353485.353487. Disponível em: <https://doi.org/10.1145/353485.353487>. Acesso em: 20 abr. 2025.

HOLMQVIST, K. et al. **Eye tracking: A comprehensive guide to methods and measures**. Oxford: Oxford University Press, 2011.

HONG, S. et al. **Metagpt: Meta programming for multi-agent collaborative framework**. 2023. arXiv:2308.00352. Disponível em: <https://arxiv.org/abs/2308.00352>. Acesso em: 20 abr. 2025.

HOULSBY, N. et al. **Parameter-Efficient Transfer Learning for NLP**. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 36., 2019, Long Beach. *Proceedings* [...]. Cambridge: PMLR, 2019. p. 2790–2799. (Proceedings of Machine Learning Research, v. 97). Disponível em: <https://proceedings.mlr.press/v97/houlsby19a.html>. Acesso em: 20 abr. 2025.

HOWARD, J.; RUDER, S. **Universal Language Model Fine-tuning for Text Classification**. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 56., 2018, Melbourne. *Proceedings* [...], Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018. p. 328–339. DOI: 10.18653/v1/P18-1031. Disponível em: <https://doi.org/10.18653/v1/P18-1031>. Acesso em: 20 abr. 2025.

HU, Yu et al. **A hypergraph-based dual-path multi-agent trajectory prediction model with topology inferring**. Engineering Applications of Artificial Intelligence, v. 142, 110799, jul. 2025. DOI: 10.1016/j.engappai.2025.110799. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0952197625007997>. Acesso em: 27 maio 2025.

HU, E. J. et al. **LoRA: Low-Rank Adaptation of Large Language Models**. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 2022. Disponível em: <https://openreview.net/forum?id=nZeVKeeFYf9>. Acesso em: 20 abr. 2025.

HUSOM, E. J. et al. **Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency**. 2025. arXiv:2504.03360. Disponível em: <https://arxiv.org/abs/2504.03360>. Acesso em: 20 abr. 2025.

IARLORI, S. et al. **An Overview of Approaches and Methods for the Cognitive Workload Estimation in Human–Machine Interaction Scenarios through Wearables Sensors**. *BioMedInformatics*, v. 4, n. 2, p. 1155-1173, 2024. DOI: 10.3390/biomedinformatics4020059. Disponível em: <https://doi.org/10.3390/biomedinformatics4020059>. Acesso em: 20 abr. 2025.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems**. Genebra: ISO, 2019.

JIA, F. et al. **Scaling Up On-Device LLMs via Active-Weight Swapping Between DRAM and Flash**. 2025. arXiv:2504.08378. Disponível em: <https://arxiv.org/abs/2504.08378>. Acesso em: 20 abr. 2025.

JIANG, A. Q. et al. **Mistral 7B**. 2023. arXiv:2310.06825. Disponível em: <https://arxiv.org/abs/2310.06825>. Acesso em: 20 abr. 2025.

JIANG, Peiling; XIA, Haijun. **Orca: Browsing at Scale Through User-Driven and AI-Facilitated Orchestration Across Malleable Webpages**. 2025. Disponível em: <https://arxiv.org/html/2505.22831v1>. Acesso em: 2 jun. 2025.

JINDAL, A. K.; RAJPOOT, P. K.; PARIKH, A. **Birbal: An efficient 7b instruct-model fine-tuned with curated datasets**. 2024. arXiv:2401.02869. Disponível em: <https://arxiv.org/abs/2401.02869>. Acesso em: 20 abr. 2025.

JOBIN, A.; IENCA, M.; VAYENA, E. **The global landscape of AI ethics guidelines**. *Nature Machine Intelligence*, v. 1, n. 9, p. 389-399, 2019. DOI: 10.1038/s42256-019-0088-2. Disponível em: <https://doi.org/10.1038/s42256-019-0088-2>. Acesso em: 20 abr. 2025.

JOUPPI, N. P. et al. **In-datacenter performance analysis of a tensor processing unit**. *In*: ANNUAL INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE (ISCA), 44., 2017, Toronto. *Proceedings* [...]. New York: ACM, 2017. p. 1–12. DOI: 10.1145/3079856.3079872. Disponível em: <https://doi.org/10.1145/3079856.3079872>. Acesso em: 20 abr. 2025.

KAHNEMAN, Daniel. **Thinking, fast and slow**. New York: Farrar, Straus and Giroux, 2011.

KAHNEMAN, D.; BEATTY, J. **Pupil diameter and load on memory**. *Science*, v. 154, n. 3756, p. 1583-1585, 1966. DOI: 10.1126/science.154.3756.1583. Disponível em: <https://doi.org/10.1126/science.154.3756.1583>. Acesso em: 20 abr. 2025.

KALYUGA, S. et al. **The expertise reversal effect**. *Educational Psychologist*, v. 38, n. 1, p. 23-31, 2003. DOI: 10.1207/S15326985EP3801\_4. Disponível em: <https://doi.org/10.1207/S15326985EP3801_4>. Acesso em: 20 abr. 2025.

KANGO, M. et al. Deep Learning Based Brain Tumor Detection and Classification Using MRI Images. **Indian Journal of Technical Education**, v. 47, Special Issue, n. 2, p. 10-16, ago. 2024. Disponível em: <https://isteonline.org/downloadables/Spl-Issue/Spl%20issue%202%20August%202024%20for%20web%20uploading.pdf>. Acesso em: 14 maio 2025.

KAPLAN, J. et al. **Scaling laws for neural language models**. 2020. arXiv:2001.08361. Disponível em: <https://arxiv.org/abs/2001.08361>. Acesso em: 20 abr. 2025.

KAZEMITABAAR, M. et al. **CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs**. 2024. arXiv:2402.01380. Disponível em: <https://arxiv.org/abs/2402.01380>. Acesso em: 20 abr. 2025.

KEOGH, E.; KASETTY, S. **On the need for time series data mining benchmarks: a survey and empirical demonstration**. *Data Mining and Knowledge Discovery*, v. 7, n. 4, p. 349-371, 2003. DOI: 10.1023/A:1024988512476. Disponível em: <https://doi.org/10.1023/A:1024988512476>. Acesso em: 20 abr. 2025.

KHRAMOV, Nikita et al. RocqStar: **Leveraging Similarity-driven Retrieval and Agentic Systems for Rocq generation**. 2025. Disponível em: <https://www.arxiv.org/pdf/2505.22846>. Acesso em: 2 jun. 2025.

KIM, M.; ORSO, A.; SINHA, S. **LlamaRestTest: Effective REST API Testing with Small Language Models**. 2025. arXiv:2501.08598. Disponível em: <https://arxiv.org/abs/2501.08598>. Acesso em: 20 abr. 2025.

KIRSCHNER, P. A.; SWELLER, J.; CLARK, R. E. **Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching**. *Educational Psychologist*, v. 41, n. 2, p. 75-86, 2006. DOI: 10.1207/s15326985ep4102\_1. Disponível em: <https://doi.org/10.1207/s15326985ep4102_1>. Acesso em: 20 abr. 2025.

KLIMESCH, W. **EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis**. *Brain Research Reviews*, v. 29, n. 2-3, p. 169-195, 1999. DOI: 10.1016/s0165-0173(98)00056-3. Disponível em: <https://doi.org/10.1016/s0165-0173(98)00056-3>. Acesso em: 20 abr. 2025.

KOTHE, C. **LabStreamingLayer**. GitHub Repository, 2014. Disponível em: <https://github.com/sccn/labstreaminglayer>. Acesso em: 20 abr. 2025.

KRAMER, A. F. **Physiological metrics of mental workload: A review of recent progress**. *In*: DAMOS, D. L. (Ed.). **Multiple-task performance**. London: Taylor & Francis, 1991. p. 279-328.

KRIGOLSON, O. E. et al. **Identifying the ERP components associated with reward processing**. *Neuropsychologia*, v. 96, p. 1-14, 2017. DOI: 10.1016/j.neuropsychologia.2017.01.001. Disponível em: <https://doi.org/10.1016/j.neuropsychologia.2017.01.001>. Acesso em: 20 abr. 2025.

KRISHNAN, Rajesh. Model Context Protocol: standardizing context management for multi-agent AI systems. In: **INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS**, 24., 2025, London. Proceedings [...]. London: AAMAS Press, 2025. p. 1156-1164. Acesso em: 27 maio 2025.

KUMAR, K. et al. **LLM Post-Training: A Deep Dive into Reasoning Large Language Models**. 2025. arXiv:2502.21321. Disponível em: <https://arxiv.org/abs/2502.21321>. Acesso em: 20 abr. 2025.

KWON, S. et al. **A Novel Personalized Healthcare System for Detecting Atrial Fibrillation Based on a Wearable Electrocardiogram Device**. *Sensors*, v. 23, n. 1, p. 134, 2023. DOI: 10.3390/s23010134. Disponível em: <https://doi.org/10.3390/s23010134>. Acesso em: 20 abr. 2025.

LABAN, Philippe et al. Are LLMs getting lost in multi-turn conversations? A study of coherence and reliability degradation. **Journal of Artificial Intelligence Research**, v. 89, p. 245-267, 2025. Disponível em: <https://www.jair.org/index.php/jair/article/view/14892>. Acesso em: 27 maio 2025.

LACOSTE, A. et al. **Quantifying the Carbon Emissions of Machine Learning**. *In*: WORKSHOP ON TACKLING CLIMATE CHANGE WITH MACHINE LEARNING AT NEURIPS 2019, 2020. arXiv:1910.09700. Disponível em: <https://arxiv.org/abs/1910.09700>. Acesso em: 20 abr. 2025.

LAMBIOTTE, R.; ROSVALL, M.; SCHOLTES, I. **From networks to optimal higher-order models of complex systems**. *Nature Physics*, v. 15, n. 4, p. 313-320, 2019. DOI: 10.1038/s41567-019-0459-y. Disponível em: <https://doi.org/10.1038/s41567-019-0459-y>. Acesso em: 20 abr. 2025.

LANGCHAIN. **LangChain Documentation**. 2023. Disponível em: <https://python.langchain.com/docs/get_started/introduction>. Acesso em: 20 abr. 2025.

LANGCHAIN. **LangGraph Documentation**. 2024. Disponível em: <https://langchain-ai.github.io/langgraph/>. Acesso em: 20 abr. 2025.

LESTER, B.; AL-RFOU, R.; CONSTANT, N. **The Power of Scale for Parameter-Efficient Prompt Tuning**. 2021. arXiv:2104.08691. Disponível em: <https://arxiv.org/abs/2104.08691>. Acesso em: 20 abr. 2025.

LEWIS, P. et al. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. *Advances in Neural Information Processing Systems*, v. 33, p. 9459-9474, 2020. Disponível em: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>. Acesso em: 20 abr. 2025.

LI, K.; HOPKINS, A.; BAUCKHAGE, C. **LoftQ: Enhanced Initialization for Low-Rank Adaptation**. 2023. arXiv:2310.05678. Disponível em: <https://arxiv.org/abs/2310.05678>. Acesso em: 20 abr. 2025.

LI, X. et al. **SVDQuant: A Novel SVD-Based Quantization Scheme for Large Language Models**. 2024. arXiv:2402.14831. Disponível em: <https://arxiv.org/abs/2402.14831>. Acesso em: 20 abr. 2025.

LIALIN, V.; DESHWAL, A.; RUMSHISKY, A. **Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning**. 2023. arXiv:2303.15647. Disponível em: <https://arxiv.org/abs/2303.15647>. Acesso em: 20 abr. 2025.

LIN, J. et al. **AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration**. 2024. arXiv:2306.00978. Disponível em: <https://arxiv.org/abs/2306.00978>. Acesso em: 20 abr. 2025.

LIN, Kevin et al. **Sleep-time Compute: Beyond Inference Scaling at Test-time**. arXiv preprint arXiv:2504.13171, 17 abr. 2025. Disponível em: <https://arxiv.org/html/2504.13171v1>. Acesso em: 27 maio 2025.

LIU, Y. et al. **AgentBench: Evaluating LLMs as Agents**. 2025. arXiv:2308.03688. Disponível em: <https://arxiv.org/abs/2308.03688>. Acesso em: 20 abr. 2025.

LUO, J. et al. **Large Language Model Agent: A Survey on Methodology, Applications and Challenges**. 2025a. arXiv:2503.21460. Disponível em: <https://arxiv.org/abs/2503.21460>. Acesso em: 20 abr. 2025.

LUO, R. et al. **AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework**. 2025b. arXiv:2308.08155. Disponível em: <https://arxiv.org/abs/2308.08155>. Acesso em: 20 abr. 2025.

LUO, Shirong et al. **Agentic AI: A Survey and Outlook**. arXiv preprint arXiv:2502.08252, 2025c. Disponível em: <https://arxiv.org/abs/2502.08252>. Acesso em: 27 maio 2025.

LV, K. et al. **DuoDecoding: Hardware-aware Heterogeneous Speculative Decoding with Dynamic Multi-Sequence Drafting**. 2025. arXiv:2503.00784. Disponível em: <https://arxiv.org/abs/2503.00784>. Acesso em: 20 abr. 2025.

MA, S. et al. **The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits**. 2024a. arXiv:2402.17764. Disponível em: <https://arxiv.org/abs/2402.17764>. Acesso em: 20 abr. 2025.

MA, X. et al. **Determining Cognitive Workload Using Physiological Measurements: Pupillometry and Heart-Rate Variability**. *Sensors*, v. 24, n. 6, p. 2010, 2024b. DOI: 10.3390/s24062010. Disponível em: <https://doi.org/10.3390/s24062010>. Acesso em: 20 abr. 2025.

MA, Hao et al. **Coevolving with the Other You: Fine-Tuning LLM with Sequential Cooperative Multi-Agent Reinforcement Learning**. arXiv preprint arXiv:2410.06101, 22 fev. 2025c. Disponível em: <https://arxiv.org/html/2410.06101v2>. Acesso em: 27 maio 2025.

MABROK, Mohamed A. et al. **Human models in human-in-the-loop control systems**. Journal of Intelligent & Fuzzy Systems, v. 38, p. 2611-2622, 2020. DOI: 10.3233/JIFS-179548. Disponível em: <https://www.researchgate.net/profile/Mohamed-Mabrok-2/publication/337880739_Human_models_in_human-in-the-loop_control_systems/links/5e6ac216299bf12e23c040fb/Human-models-in-human-in-the-loop-control-systems.pdf>. Acesso em: 2 jun. 2025.

MAKOWSKI, D. et al. **NeuroKit2: A Python toolbox for neurophysiological signal processing**. *Behavior Research Methods*, v. 53, n. 4, p. 1689-1696, 2021. DOI: 10.3758/s13428-020-01516-y. Disponível em: <https://doi.org/10.3758/s13428-020-01516-y>. Acesso em: 20 abr. 2025.

MARJANOVIĆ, S. V. et al. **DeepSeek-R1 Thoughtology: Let’s about LLM reasoning**. 2025. arXiv:2504.07128. Disponível em: <https://arxiv.org/abs/2504.07128>. Acesso em: 20 abr. 2025.

MAYER, R. E.; MORENO, R. **Nine ways to reduce cognitive load in multimedia learning**. *Educational Psychologist*, v. 38, n. 1, p. 43-52, 2003. DOI: 10.1207/S15326985EP3801\_6. Disponível em: <https://doi.org/10.1207/S15326985EP3801_6>. Acesso em: 20 abr. 2025.

MICROSOFT. **Introducing NLWeb: Bringing conversational interfaces directly to the web**. Microsoft News, 19 maio 2025. Disponível em: <https://news.microsoft.com/source/features/company-news/introducing-nlweb-bringing-conversational-interfaces-directly-to-the-web/>. Acesso em: 27 maio 2025.

MISTRAL. **Medium is the new large**. Mistral AI Blog, 7 maio 2025. Disponível em: <https://mistral.ai/news/mistral-medium-3>. Acesso em: 27 maio 2025.

MWAMBA, E.; NKOSI, F. **Human-Centric AI: Understanding and Enhancing Collaboration between Humans and Intelligent Systems**. *Journal of Algorithm Asynchronous (Algrasy)*, v. 1, n. 1, p. 33-40, 2023. Disponível em: <https://hasmed.org/index.php/jourasy/article/view/49>. Acesso em: 20 abr. 2025.

NAGEL, M. et al. **A White Paper on Neural Network Quantization**. 2021. arXiv:2106.08295. Disponível em: <https://arxiv.org/abs/2106.08295>. Acesso em: 20 abr. 2025.

NAKAGAWA, S.; SCHIELZETH, H. **A general and simple method for obtaining R² from generalized linear mixed-effects models**. *Methods in Ecology and Evolution*, v. 4, n. 2, p. 133-142, 2013. DOI: 10.1111/j.2041-210x.2012.00261.x. Disponível em: <https://doi.org/10.1111/j.2041-210x.2012.00261.x>. Acesso em: 20 abr. 2025.

NGUYEN, H. C.; MAMITSUKA, H. **Learning on hypergraphs with sparsity**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 44, n. 10, p. 7113-7127, 2022. DOI: 10.1109/TPAMI.2020.2974746. Disponível em: <https://doi.org/10.1109/TPAMI.2020.2974746>. Acesso em: 20 abr. 2025.

NI, Jiliang et al. **From Large to Super-Tiny: End-to-End Optimization for Cost-Efficient LLMs**. 2025. Preprint. arXiv:2504.13471v1 [cs.CL]. Disponível em: <https://arxiv.org/html/2504.13471v1>. Acesso em: 24 abr. 2025.

NIELSEN, J. **Usability Engineering**. Boston: Academic Press, 1993.

NORMAN, D. A. **The psychology of everyday things**. New York: Basic Books, 1988.

OPENAI. **A practical guide to building agents**. OpenAI Business Guides and Resources, 2025. Disponível em: <https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf>. Acesso em: 20 abr. 2025.

OROVAS, C. et al. **EEG in Education: A Scoping Review of Hardware, Software, and Methodological Aspects**. *Brain Sciences*, v. 15, n. 2, p. 203, 2024. DOI: 10.3390/brainsci15020203. Disponível em: <https://doi.org/10.3390/brainsci15020203>. Acesso em: 20 abr. 2025.

OSBORNE, J. W.; OVERBAY, A. **The power of outliers (and why researchers should ALWAYS check for them)**. *Practical Assessment, Research, and Evaluation*, v. 9, n. 6, 2004. Disponível em: <https://scholarworks.umass.edu/pare/vol9/iss1/6/>. Acesso em: 20 abr. 2025.

OUYANG, L. et al. **Training language models to follow instructions with human feedback**. 2022. arXiv:2203.02155. Disponível em: <https://arxiv.org/abs/2203.02155>. Acesso em: 20 abr. 2025.

OZDEMIR, Serkan; GATTI, Marco. Social touch modeling in multi-agent reinforcement learning: regulating internal states through arousal mechanisms. **IEEE Transactions on Cognitive and Developmental Systems**, v. 17, n. 3, p. 892-905, 2025. DOI: 10.1109/TCDS.2024.3487291. Acesso em: 27 maio 2025.

PAAS, F. **Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach**. *Journal of Educational Psychology*, v. 84, n. 4, p. 429–434, 1992. DOI: 10.1037/0022-0663.84.4.429. Disponível em: <https://doi.org/10.1037/0022-0663.84.4.429>. Acesso em: 20 abr. 2025.

PAAS, F. et al. **Cognitive load measurement as a means to advance cognitive load theory**. *Educational Psychologist*, v. 38, n. 1, p. 63-71, 2003. DOI: 10.1207/S15326985EP3801\_8. Disponível em: <https://doi.org/10.1207/S15326985EP3801_8>. Acesso em: 20 abr. 2025.

PAAS, F.; VAN MERRIËNBOER, J. J. G. **Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach**. *Journal of Educational Psychology*, v. 86, n. 1, p. 122–133, 1994. DOI: 10.1037/0022-0663.86.1.122. Disponível em: <https://doi.org/10.1037/0022-0663.86.1.122>. Acesso em: 20 abr. 2025.

PAHUNE, Saurabh; AKHTAR, Zahid. **Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models**. Information, v. 16, n. 2, 87, 2025. DOI: 10.3390/info16020087. Disponível em: <https://www.mdpi.com/2078-2489/16/2/87>. Acesso em: 24 abr. 2025.

PARK, J. S. et al. **Generative Agents: Interactive Simulacra of Human Behavior**. 2023. arXiv:2304.03442. Disponível em: <https://arxiv.org/abs/2304.03442>. Acesso em: 20 abr. 2025.

PAULOSKI, J. Gregory et al. Academy: a middleware platform for autonomous agent deployment in federated research infrastructures. **Future Generation Computer Systems**, v. 154, p. 78-92, maio 2025. DOI: 10.1016/j.future.2024.12.003. Acesso em: 27 maio 2025.

PENA-PENA, K. et al. **Learning hypergraphs tensor representations from data via t-hgsp**. *IEEE Transactions on Signal and Information Processing over Networks*, v. 10, p. 17–31, 2024. DOI: 10.1109/TSIPN.2023.3346205. Disponível em: <https://doi.org/10.1109/TSIPN.2023.3346205>. Acesso em: 20 abr. 2025.

PENG, Y. et al. **GraphRAG: A Retrieval-Augmented Generation Framework for Large Language Models on Graph-Structured Data**. 2024. arXiv:2404.17146. Disponível em: <https://arxiv.org/abs/2404.17146>. Acesso em: 20 abr. 2025.

POPE, A. T.; BOGART, E. H.; BARTOLOME, D. S. **Biocybernetic system evaluates indices of operator engagement in automated task**. *Biological Psychology*, v. 40, n. 1-2, p. 187-195, 1995. DOI: 10.1016/0301-0511(95)05126-1. Disponível em: <https://doi.org/10.1016/0301-0511(95)05116-3>. Acesso em: 20 abr. 2025.

PRIGOGINE, Ilya. **The end of certainty**. New York: Free Press, 1997.

PROMPIENGCHAI, Sapolnach; NARREDDY, Charith; JOORDENS, Steve. **A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI**. 2025. Disponível em: <https://www.arxiv.org/pdf/2505.23405>. Acesso em: 2 jun. 2025.

PU, Xiao et al. **ThoughtTerminator: Benchmarking, Calibrating, and Mitigating Overthinking in Reasoning Models**. 2025. Preprint. arXiv:2504.13367v1 [cs.CL]. Disponível em: <https://arxiv.org/html/2504.13367v1>. Acesso em: 24 abr. 2025.

RADIVOJEVIC, Kristina et al. **Public Discourse Sandbox: Facilitating Human and AI Digital Communication Research**. 2025. Disponível em: <https://arxiv.org/html/2505.21604>. Acesso em: 2 jun. 2025.

RADOSEVICH, B.; HALLORAN, J. T. **MCP Safety Audit: LLMs with the Model Context Protocol Allow Major Security Exploits**. 2025. arXiv:2504.03767. Disponível em: <https://arxiv.org/abs/2504.03767>. Acesso em: 20 abr. 2025.

RAFFEL, C. et al. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. *Journal of Machine Learning Research*, v. 21, n. 140, p. 1-67, 2020. Disponível em: <https://jmlr.org/papers/v21/20-074.html>. Acesso em: 20 abr. 2025.

RANALDI, L.; HADDOW, B.; BIRCH, A. **Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task**. 2025. arXiv:2504.03616. Disponível em: <https://arxiv.org/abs/2504.03616>. Acesso em: 20 abr. 2025.

RAYNER, K. **Eye movements in reading and information processing: 20 years of research**. *Psychological Bulletin*, v. 124, n. 3, p. 372-422, 1998. DOI: 10.1037/0033-2909.124.3.372. Disponível em: <https://doi.org/10.1037/0033-2909.124.3.372>. Acesso em: 20 abr. 2025.

RODRIGUEZ-RAMIREZ, A.; VERGARA VILLEGAS, O. O.; NANDAYAPA, M.; GARCIA-LUNA, F.; GUEVARA NERI, M. C. A Diamond Approach to Develop Virtual Object Interaction: Fusing Augmented Reality and Kinesthetic Haptics. **Crossmodal Technologies and Interaction**, v. 9, n. 2, art. 15, 2025. Disponível em: <https://www.mdpi.com/2414-4088/9/2/15>. Acesso em: 14 maio 2025.

RUBIN, D. B. **Multiple imputation for nonresponse in surveys**. New York: Wiley, 1987.

RUZZELLI, A. G.; NICOLAS, C. O.; MCGIBNEY, A.; RUZELLI, F. **Real-Time Recognition and Profiling of Appliances through High-Frequency Current Monitoring**. In: SUSTAINABLE INTERNET AND ICT FOR SUSTAINABILITY (SustainIT), 2010 IEEE International Symposium on. [S.l.]: IEEE, 2010. p. 1-6. Disponível em: <https://core.ac.uk/download/197525877.pdf>. Acesso em: 14 maio 2025.

SALVUCCI, D. D.; TAATGEN, N. A. **Threaded cognition: An integrated theory of concurrent multitasking**. *Psychological Review*, v. 115, n. 1, p. 101-130, 2008. DOI: 10.1037/0033-295X.115.1.101. Disponível em: <https://doi.org/10.1037/0033-295X.115.1.101>. Acesso em: 20 abr. 2025.

SAROUFIM, M. et al. **NeurIPS 2023 LLM Efficiency Fine-tuning Competition**. 2025. arXiv:2503.13507. Disponível em: <https://arxiv.org/abs/2503.13507>. Acesso em: 20 abr. 2025.

SCHWARTZ, R. et al. **Green AI**. *Communications of the ACM*, v. 63, n. 12, p. 54-63, 2020. DOI: 10.1145/3381831. Disponível em: <https://doi.org/10.1145/3381831>. Acesso em: 20 abr. 2025.

SHAO, W. et al. **OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models**. 2024. arXiv:2308.13137. Disponível em: <https://arxiv.org/abs/2308.13137>. Acesso em: 20 abr. 2025.

SHARMA, N.; KUMAR, P.; LI, Y. **OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models**. 2024. arXiv:2412.15235. Disponível em: <https://arxiv.org/abs/2412.15235>. Acesso em: 20 abr. 2025.

SHIFFRIN, R. M.; SCHNEIDER, W. **Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory**. *Psychological Review*, v. 84, n. 2, p. 127–190, 1977. DOI: 10.1037/0033-295X.84.2.127. Disponível em: <https://doi.org/10.1037/0033-295X.84.2.127>. Acesso em: 20 abr. 2025.

SIMON, Herbert. **Models of Man**. New York: John Wiley & Sons, 1957.

SOSA, M. E.; EPPINGER, S. D.; ROWLES, C. M. **The misalignment of product architecture and organizational structure in complex product development**. *Management Science*, v. 50, n. 12, p. 1674-1689, 2004. DOI: 10.1287/mnsc.1040.0289. Disponível em: <https://doi.org/10.1287/mnsc.1040.0289>. Acesso em: 20 abr. 2025.

SPIESS, Claudio et al. **AutoPDL: Automatic Prompt Optimization for LLM Agents**. 2025. Preprint. arXiv:2504.04365v1 [cs.LG]. Disponível em: <https://arxiv.org/pdf/2504.04365>. Acesso em: 24 abr. 2025.

STIENNON, N. et al. **Learning to summarize from human feedback**. 2020. arXiv:2009.01325. Disponível em: <https://arxiv.org/abs/2009.01325>. Acesso em: 20 abr. 2025.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. **Energy and Policy Considerations for Deep Learning in NLP**. 2019. arXiv:1906.02243. Disponível em: <https://arxiv.org/abs/1906.02243>. Acesso em: 20 abr. 2025.

SWELLER, J. **Cognitive Architecture and Instructional Design**. *Educational Psychology Review*, v. 10, n. 3, p. 251-296, 1998. DOI: 10.1023/A:1022193728205. Disponível em: <https://doi.org/10.1023/A:1022193728205>. Acesso em: 20 abr. 2025.

SWELLER, J. **Cognitive load during problem solving: Effects on learning**. *Cognitive Science*, v. 12, n. 2, p. 257-285, 1988. DOI: 10.1207/s15516709cog1202\_4. Disponível em: <https://doi.org/10.1207/s15516709cog1202_4>. Acesso em: 20 abr. 2025.

SWELLER, J. **Cognitive load theory and individual differences**. *Learning and Individual Differences*, v. 109, p. 102423, 2024. DOI: 10.1016/j.lindif.2024.102423. Disponível em: <https://doi.org/10.1016/j.lindif.2024.102423>. Acesso em: 20 abr. 2025.

SWELLER, J. **Cognitive load theory, learning difficulty, and instructional design**. *Learning and Instruction*, v. 4, n. 4, p. 295-312, 1994. DOI: 10.1016/0959-4752(94)90003-5. Disponível em: <https://doi.org/10.1016/0959-4752(94)90003-5>. Acesso em: 20 abr. 2025.

SWELLER, J. **Element interactivity and intrinsic, extraneous, and germane cognitive load**. *Educational Psychology Review*, v. 22, n. 2, p. 123-138, 2010. DOI: 10.1007/s10648-010-9128-5. Disponível em: <https://doi.org/10.1007/s10648-010-9128-5>. Acesso em: 20 abr. 2025.

SWELLER, J.; CHANDLER, P. **Why some material is difficult to learn**. *Cognition and Instruction*, v. 12, n. 3, p. 185-233, 1994. DOI: 10.1207/s1532690xci1203\_1. Disponível em: <https://doi.org/10.1207/s1532690xci1203_1>. Acesso em: 20 abr. 2025.

SWELLER, J.; VAN MERRIENBOER, J. J. G.; PAAS, F. **Cognitive Architecture and Instructional Design**. *Educational Psychology Review*, v. 10, n. 3, p. 251-296, 1998. DOI: 10.1023/A:1022193728205. Disponível em: <https://doi.org/10.1023/A:1022193728205>. Acesso em: 20 abr. 2025.

SWELLER, John. **Cognitive Load Theory**. In: ROSS, B. H. (Ed.). Psychology of Learning and Motivation. v. 55. San Diego: Academic Press, 2011. p. 37-76.

TANG, B.; CHEN, S.; DONG, X. **Hypergraph Structure Inference From Data Under Smoothness Prior**. 2023. arXiv:2308.14172. Disponível em: <https://arxiv.org/abs/2308.14172>. Acesso em: 20 abr. 2025.

TANTITHAMTHAVORN, Chakkrit (Kla) et al. **MLOps, LLMOps, FMOps, and Beyond**. IEEE Software, v. 42, n. 1, p. 26-32, jan./fev. 2025. DOI: 10.1109/MS.2024.3477014.

TOUVRON, H. et al. **Llama 2: Open Foundation and Fine-Tuned Chat Models**. 2023. arXiv:2307.09288. Disponível em: <https://arxiv.org/abs/2307.09288>. Acesso em: 20 abr. 2025.

TRAUNMUELLER, P. et al. **Wearable Healthcare Devices for Monitoring Stress and Attention Level in Workplace Environments**. 2024. arXiv:2406.05813. Disponível em: <https://arxiv.org/abs/2406.05813>. Acesso em: 20 abr. 2025.

TRUONG, Q.-T. et al. **Towards Automated Testing of LLM-based Applications: A Survey**. 2023. arXiv:2311.18119. Disponível em: <https://arxiv.org/abs/2311.18119>. Acesso em: 20 abr. 2025.

VALIPOUR, H. et al. **DyLoRA: Parameter-Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation**. 2022. arXiv:2210.07558. Disponível em: <https://arxiv.org/abs/2210.07558>. Acesso em: 20 abr. 2025.

VAN DER AALST, W. M. P. **Process mining: data science in action**. 2. ed. Berlin: Springer, 2016.

VAN MERRIËNBOER, J. J. G.; SWELLER, J. **Cognitive load theory and complex learning: Recent developments and future directions**. *Educational Psychology Review*, v. 17, n. 2, p. 147-177, 2005. DOI: 10.1007/s10648-005-3951-0. Disponível em: <https://doi.org/10.1007/s10648-005-3951-0>. Acesso em: 20 abr. 2025.

VAN SOMEREN, M. W.; BARNARD, Y. F.; SANDBERG, J. A. C. **The think aloud method: A practical guide to modelling cognitive processes**. London: Academic Press, 1994.

WANG, H. et al. **Leveraging Reasoning Model Answers to Enhance Non-Reasoning Model Capability**. 2025a. arXiv:2504.09639. Disponível em: <https://arxiv.org/abs/2504.09639>. Acesso em: 20 abr. 2025.

WANG, W. et al. **FedPETuning: Federated Parameter-Efficient Fine-Tuning of Large Language Models**. 2025b. arXiv:2311.18641. Disponível em: <https://arxiv.org/abs/2311.18641>. Acesso em: 20 abr. 2025.

WANG, Junlin et al. **Think Deep, Think Fast: Investigating Efficiency of Verifierfree Inference-time-scaling Methods**. 2025c. Preprint. arXiv:2504.14047v1 [cs.AI]. Disponível em: <https://arxiv.org/pdf/2504.14047>. Acesso em: 24 abr. 2025.

WANG, Hongru et al. **OTC: Optimal Tool Calls via Reinforcement Learning**. 2025d. Preprint. arXiv:2504.14870v1 [cs.AI]. Disponível em: <https://arxiv.org/html/2504.14870v1>. Acesso em: 24 abr. 2025.

WEI, J. et al. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. 2022a. arXiv:2201.11903. Disponível em: <https://arxiv.org/abs/2201.11903>. Acesso em: 20 abr. 2025.

WEI, J. et al. **Emergent Abilities of Large Language Models**. 2022b. arXiv:2206.07682. Disponível em: <https://arxiv.org/abs/2206.07682>. Acesso em: 20 abr. 2025.

WESKE, M. **Business process management: concepts, languages, architectures**. 3. ed. Berlin: Springer, 2019.

WEYSSOW, M. et al. **Exploring parameter-efficient fine-tuning techniques for code generation with large language models**. 2023. arXiv:2308.10462. Disponível em: <https://arxiv.org/abs/2308.10462>. Acesso em: 20 abr. 2025.

WIJAYA, S. et al. **ReadMe.LLM: A Framework to Help LLMs Understand Your Library**. 2025. arXiv:2504.09798. Disponível em: <https://arxiv.org/abs/2504.09798>. Acesso em: 20 abr. 2025.

WOOD, D.; BRUNER, J. S.; ROSS, G. **The role of tutoring in problem solving**. *Journal of Child Psychology and Psychiatry*, v. 17, n. 2, p. 89-100, 1976. DOI: 10.1111/j.1469-7610.1976.tb00381.x. Disponível em: <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>. Acesso em: 20 abr. 2025.

WOOD, S. N. **Generalized additive models: an introduction with R**. 2. ed. Boca Raton: CRC press, 2017.

WU, C. et al. **Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models**. 2023. arXiv:2303.04671. Disponível em: <https://arxiv.org/abs/2303.04671>. Acesso em: 20 abr. 2025.

WU, T. et al. **AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework**. 2023. arXiv:2308.08155. Disponível em: <https://arxiv.org/abs/2308.08155>. Acesso em: 20 abr. 2025.

XIA, Shijie et al. **Unlocking Deep Thinking in Language Models: Cognition Engineering through Inference Time Scaling and Reinforcement Learning**. 2025. Preprint. arXiv:2504.13828v1 [cs.CL]. Disponível em: <https://arxiv.org/html/2504.13828v1>. Acesso em: 24 abr. 2025.

XIAO, G. et al. **SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models**. 2023. arXiv:2211.10438. Disponível em: <https://arxiv.org/abs/2211.10438>. Acesso em: 20 abr. 2025.

XU, W. **Human-Centered AI**. Boca Raton: CRC Press, 2019.

XU, Y. et al. **QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models**. 2023. arXiv:2309.14717. Disponível em: <https://arxiv.org/abs/2309.14717>. Acesso em: 20 abr. 2025.

YAO, S. et al. **ReAct: Synergizing Reasoning and Acting in Language Models**. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 2023. Disponível em: <https://openreview.net/forum?id=67WPM9hC5li>. Acesso em: 20 abr. 2025.

YERKES, R. M.; DODSON, J. D. **The relation of strength of stimulus to rapidity of habit‐formation**. *Journal of Comparative Neurology and Psychology*, v. 18, n. 5, p. 459-482, 1908. DOI: 10.1002/cne.920180503. Disponível em: <https://doi.org/10.1002/cne.920180503>. Acesso em: 20 abr. 2025.

YU, Y. et al. **Cognitive Load Prediction From Crossmodal Physiological Signals using Multiview Learning**. *IEEE Journal of Biomedical and Health Informatics*, v. 28, n. 1, p. 486-496, 2024. DOI: 10.1109/JBHI.2023.3346205. Disponível em: <https://doi.org/10.1109/JBHI.2023.3346205>. Acesso em: 20 abr. 2025.

ZHANG, D. et al. **ReSTMCTS: LLM self-training via process reward guided tree search**\*. *Advances in Neural Information Processing Systems*, v. 37, p. 64735–64772, 2024. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2024/hash/76ec4dc30e9faaf0e4b6093eaa377218-Abstract-Conference.html>. Acesso em: 20 abr. 2025.

ZHANG, J.; NORMAN, D. A. **Representations in distributed cognitive tasks**. *Cognitive Science*, v. 18, n. 1, p. 87-122, 1994. DOI: 10.1207/s15516709cog1801\_3. Disponível em: <https://doi.org/10.1207/s15516709cog1801_3>. Acesso em: 20 abr. 2025.

ZHANG, T. et al. **LLM-Orchestrator: Orchestrating Collaborative Large Language Models for Automated Software Engineering**. 2025a. arXiv:2402.16831. Disponível em: <https://arxiv.org/abs/2402.16831>. Acesso em: 20 abr. 2025.

ZHANG, Lingzhe et al. **XRAGLog: A Resource-Efficient and Context-Aware Log-Based Anomaly Detection Method Using Retrieval-Augmented Generation**. 2025b. Preprint/Submissão de Conferência. Disponível em: <https://openreview.net/pdf?id=8gv7CXuXQ3>. Acesso em: 24 abr. 2025.

ZHAO, P.; YAUN, X. **GANQ: GPU-Adaptive Non-Uniform Quantization for Large Language Models**. 2025a. arXiv:2501.12956. Disponível em: <https://arxiv.org/abs/2501.12956>. Acesso em: 20 abr. 2025.

ZHAO, Baining et al. **Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning in Foundation Models via Reinforcement Learning**. 2025b. Preprint. arXiv:2504.12680v1 [cs.AI]. Disponível em: <https://arxiv.org/html/2504.12680v1>. Acesso em: 24 abr. 2025.

ZHUGE, M. et al. **Mindstorms in Natural Language: A Survey of LLM-based Agents**. 2024. arXiv:2311.01398. Disponível em: <https://arxiv.org/abs/2311.01398>. Acesso em: 20 abr. 2025.

ZIMMERMAN, B. J. **Becoming a self-regulated learner: An overview**. *Theory Into Practice*, v. 41, n. 2, p. 64-70, 2002. DOI: 10.1207/s15430421tip4102\_2. Disponível em: <https://doi.org/10.1207/s15430421tip4102_2>. Acesso em: 20 abr. 2025.

ZOU, L. et al. **GraphRAG: A Survey on Retrieval-Augmented Generation for Graph-Structured Data**. 2025. arXiv:2504.08967. Disponível em: <https://arxiv.org/abs/2504.08967>. Acesso em: 20 abr. 2025.

**APÊNDICE I -** NASA-TLX Adaptado

### PARTE I: Escalas de Avaliação

*Para cada dimensão, marque um "X" na escala de 21 pontos que melhor representa sua experiência durante a tarefa com os modelos GGUF.*

#### 1. DEMANDA MENTAL

**Quanto esforço mental e perceptual foi necessário para gerenciar o ajuste fino QLoRA e orquestrar os agentes LLM?**

Muito Baixa |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito Alta

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

#### 2. DEMANDA FÍSICA

**Quanto esforço físico foi necessário para interagir com a interface CrossDebate e sensores de monitoramento?**

Muito Baixa |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito Alta

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

#### 3. DEMANDA TEMPORAL

**Quanto você sentiu pressão de tempo ao configurar parâmetros, monitorar métricas CL/CompL e gerenciar os fluxos multiagente?**

Muito Baixa |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito Alta

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

#### 4. DESEMPENHO PERCEBIDO

**Quão bem-sucedido você considera ter sido em completar o ajuste fino dos modelos GGUF e orquestrar os agentes?**

Perfeito |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Falhou

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

#### 5. ESFORÇO

**Quanto você teve que trabalhar (mental e fisicamente) para alcançar seu nível de desempenho na tarefa?**

Muito Baixo |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito Alto

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

#### 6. FRUSTRAÇÃO

**Quanto você se sentiu inseguro, desencorajado, irritado, estressado ou aborrecido durante a interação com os modelos quantizados e a plataforma?**

Muito Baixa |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito Alta

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

### PARTE II: Comparações Pareadas - Pesos Relativos

*Para cada par abaixo, clique na fonte de carga de trabalho que contribuiu MAIS para a carga de trabalho durante sua tarefa com a plataforma CrossDebate.*

| **Comparação** | **Opção A** | **Opção B** |
| --- | --- | --- |
| 1 | ☐ Demanda Mental | ☐ Desempenho Percebido |
| 2 | ☐ Demanda Temporal | ☐ Demanda Mental |
| 3 | ☐ Demanda Temporal | ☐ Esforço |
| 4 | ☐ Demanda Física | ☐ Frustração |
| 5 | ☐ Desempenho Percebido | ☐ Frustração |
| 6 | ☐ Demanda Física | ☐ Demanda Temporal |
| 7 | ☐ Demanda Física | ☐ Desempenho Percebido |
| 8 | ☐ Demanda Temporal | ☐ Frustração |
| 9 | ☐ Esforço | ☐ Frustração |
| 10 | ☐ Desempenho Percebido | ☐ Esforço |
| 11 | ☐ Demanda Mental | ☐ Demanda Física |
| 12 | ☐ Demanda Mental | ☐ Esforço |
| 13 | ☐ Demanda Mental | ☐ Frustração |
| 14 | ☐ Demanda Física | ☐ Esforço |
| 15 | ☐ Demanda Temporal | ☐ Desempenho Percebido |

### PARTE III: Questões Específicas do Estudo CrossDebate

#### A. Impacto da Quantização GGUF

**Como a diferença entre modelos Q4 e Q8 afetou sua carga de trabalho?**

1. **Interpretação de Saídas Q4 vs Q8:**

Q4 muito mais difícil |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Q8 muito mais difícil

-5 -4 -3 -2 -1 0 1 2 3 4 5

1. **Confiança nas Decisões com Q4 vs Q8:**

Muito menos confiante com Q4 |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito menos confiante com Q8

-5 -4 -3 -2 -1 0 1 2 3 4 5

#### B. Complexidade Multiagente

**Como a orquestração de múltiplos agentes LLM afetou sua experiência?**

1. **Agente Único → Multiagente Sequencial → Multiagente Paralelo**

Sem mudança na carga |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Aumento dramático na carga

0 1 2 3 4 5 6 7 8 9 10

1. **Maior desafio na orquestração multiagente foi:**

* ☐ Configurar dependências entre agentes
* ☐ Monitorar execução distribuída
* ☐ Integrar resultados parciais
* ☐ Depurar falhas de coordenação
* ☐ Gerenciar recursos computacionais
* ☐ Outro: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#### C. Interface e Feedback da Plataforma

**Quão úteis foram as visualizações CL/CompL em tempo real para suas decisões?**

Completamente inúteis |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Extremamente úteis

0 1 2 3 4 5 6 7 8 9 10

### PARTE IV: Single Ease Question (SEQ) Adaptada

**Considerando todos os aspectos (quantização, multiagente, interface), quão mentalmente demandante foi a tarefa geral?**

Muito Fácil |\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_|\_\_\_| Muito Difícil

1 2 3 4 5 6 7

### Instruções de Aplicação

1. **Timing:** Administrar imediatamente após cada bloco experimental (Q4/Q8 × Complexidade do Fluxo)
2. **Duração:** Aproximadamente 3-4 minutos por aplicação
3. **Contextualização:** Sempre especificar qual condição experimental acabou de ser completada
4. **Dados complementares:** Correlacionar com métricas fisiológicas (EEG, pupilometria, HRV) coletadas durante a tarefa

### Cálculo do Escore NASA-TLX

1. **Escores brutos:** Converter escalas de 0-100 para cada dimensão
2. **Pesos:** Calcular com base nas comparações pareadas (0-5 pontos por dimensão)
3. **Escore ponderado:** Somar (Escore bruto × Peso) para todas as dimensões
4. **Normalização:** Dividir pela soma total dos pesos (máximo 75)

**APÊNDICE II** - Glossário de Termos Técnicos

| **Termo Técnico** | **Definição Concisa (Conforme Apresentada ou Inferida na Dissertação)** | **Seção de Introdução/Discussão Principal** |
| --- | --- | --- |
| **Paradoxo da Cognição Artificial Escalonável** | Tensão entre o aumento das capacidades dos LLMs com a escala e o aumento exponencial das demandas computacionais. | Introdução |
| **Ciclo de Feedback CL-CompL** | Interação dinâmica e bidirecional onde a CL do operador afeta e é afetada pela CompL da plataforma, mediada por sensoriamento, modelagem, visualização e ação. | Discussão (Figura 1) |
| **Engenharia da Cognição** | Campo focado na construção e otimização das capacidades de pensamento da AI, incluindo test-time scaling e RL, visando modelos mais eficientes e capazes. | Discussão |
| **Cronostasis Cognitiva** | Conceito proposto para capturar a natureza temporal da CL, que se manifesta em múltiplas escalas, indo além de métricas estáticas. | Discussão |
| **Fadiga Cognitiva Adaptativa** | Framework para entender como operadores ajustam estratégias de interação em resposta à instabilidade ou sobrecarga do sistema, informando design de interfaces. | Discussão |
| **Allostasis Cognitiva** | Processo pelo qual o sistema nervoso ajusta pontos de referência homeostáticos para manter performance sob incerteza aumentada (e.g., com modelos Q4). | Discussão |
| **Falhas Silenciosas** | Erros em LLMs onde a saída é sintaticamente correta, mas semanticamente inconsistente ou degradada, dificultando a detecção pelo usuário. | Discussão |
| **Verificação Obsessiva** | Padrão de movimentos oculares (múltiplas refixações, aumento do tempo de fixação) indicando processamento esforçado ao lidar com saídas incertas de LLMs. | Discussão |
| **Regressão Esquemática** | Fenômeno onde operadores, sob alta carga ou incerteza (e.g., com Q4), perdem acesso a esquemas de expertise e regridem a estratégias de processamento mais básicas. | Discussão |
| **Paralisia por Análise** | Estado cognitivo onde o excesso de opções de configuração (e.g., em fluxos multiagente) excede a capacidade de avaliação comparativa, levando à inação ou decisão subótima. | Discussão |
| **Inversão de Responsabilidades** | Situação em sistemas humano-IA onde o operador assume tarefas que deveriam ser automatizadas, enquanto o sistema falha em suportar tarefas genuinamente humanas. | Discussão |
| **Contratos Cognitivos** | Especificações dinâmicas em sistemas híbridos humano-IA que definem qual componente é responsável por diferentes tipos de decisão, baseado em competências e carga. | Discussão |
| **Meta-protocolo** | Nível de coordenação emergente da interação sinérgica de múltiplos protocolos de comunicação (MCP, A2A, NL Web) em sistemas multiagente. | Discussão |
| **Simbiose Cognitiva** | Estado de co-adaptação em tempo real entre sistemas cognitivos humanos e artificiais, onde fronteiras de processamento se tornam difusas e há sincronização. | Discussão |
| **Articulação Teórica Transdisciplinar** | Abordagem metodológica que integra a Cognitive Load Theory com métricas de desempenho computacional e frameworks ergonômicos (ISO 9241-210:2019) para análise holística da interação humano-IA. | Introdução/Revisão da Literatura |
| **Destilação de Raciocínio** | Processo de transferência de capacidades de resolução de problemas de modelos de raciocínio mais avançados para modelos menores através de dados de treinamento que incluem etapas de pensamento explícitas. | Revisão da Literatura |
| **Cognição Distribuída Híbrida** | Forma de processamento de informação onde capacidades cognitivas humanas e artificiais se integram para formar um sistema cognitivo unificado, especialmente evidente na orquestração de sistemas multiagente. | Revisão da Literatura |
| **Mente Estendida Computacional** | Extensão da teoria de Clark e Chalmers aplicada a sistemas de AI locais, onde ferramentas tecnológicas tornam-se componentes genuínos dos processos mentais quando integradas de forma fluida nas atividades cognitivas. | Revisão da Literatura |
| **Racionalidade Limitada Computacional** | Aplicação do conceito de Herbert Simon aos sistemas de AI, onde tanto humanos quanto sistemas artificiais "satisfazem" encontrando soluções "boas o suficiente" em vez de otimização global, exemplificado pela quantização. | Introdução/Revisão da Literatura |
| **Intervenção Mínima para Máximo Impacto** | Princípio de design exemplificado pela arquitetura LoRA, onde pequenas modificações estratégicas amplificam mudanças comportamentais significativas em sistemas complexos. | Revisão da Literatura |
| **Leis de Escala Empíricas** | Correlações observadas empiricamente entre desempenho de LLMs e fatores como número de parâmetros e volume de dados de pré-treinamento, fundamentando o paradigma de crescimento em escala. | Introdução |
| **Habilidades Emergentes** | Capacidades não explicitamente presentes em modelos menores, mas que surgem quando modelos ultrapassam certos limiares de escala, sugerindo mudanças qualitativas no processamento. | Introdução/Revisão da Literatura |
| **Test-time Scaling** | Estratégias para aproveitar recursos computacionais adicionais durante inferência para melhorar desempenho sem retreinamento, como demonstrado em técnicas de sleep-time compute. | Introdução |
| **Democratização da AI Local** | Processo de tornar LLMs avançados acessíveis em hardware de prateleira através de técnicas como quantização GGUF e runtimes otimizados, removendo barreiras de acesso. | Introdução/Revisão da Literatura |