Single Prompt vs. Debate Multiagente: Uma Análise de Telemetria de Hardware em Tarefas de Contexto Longo com Llama 3.2 2B

Luiz Alexandre M. Barros [1], **Vitor Manoel A. da Silva** [1],
Thiago Angelino dos Santos [1], **Guilherme Maia R. Viana** [1],
Bruno Gomes de B. Filho [1]

[1]Instituto Atlântico (IA)
Av. Washington Soares, 909 - Edson Queiroz, Fortaleza - Ceará
alexandre_barros@atlantico.com.br, vitor_alves@atlantico.com.br
thiago_angelino@atlantico.com.br, guilherme_viana@atlantico.com.br
bruno_gomes@atlantico.com.br

Abstract. *This paper looks into whether splitting context across multiple agents actually saves energy once you account for the coordination overhead they introduce. Using direct GPU measurements through the NVIDIA Management Library (NVML) on a dedicated H100, we ran a head-to-head comparison between monolithic processing (Single Prompt) and a Dialectical Debate architecture across 100 financial-domain tasks with context windows ranging from 24k to 32k tokens. We only count energy savings as real when the multi-agent system holds its own on quality—enforced through a strict equivalence criterion ($S_{multi} \geq S_{mono}$).*

Resumo. *Este artigo investiga se a redução da carga de contexto por agente em sistemas multiagentes é suficiente para compensar o custo energético adicional da coordenação. Utilizando telemetria de hardware via NVIDIA Management Library (NVML) em uma GPU H100 dedicada, comparamos o processamento monolítico (Single Prompt) com uma arquitetura de Debate Dialético em 100 tarefas do domínio financeiro (janelas de 24k–32k tokens). Um critério de equivalência qualitativa estrita ($S_{multi} \geq S_{mono}$) garante que economias energéticas só sejam contabilizadas quando o sistema multiagente mantém a mesma qualidade do processamento monolítico.*

1. Introdução

A adoção de modelos de linguagem em tarefas com contexto longo reabriu um trade-off clássico: qualidade de resposta versus custo computacional. Em janelas de 24k–32k tokens, benchmarks como LongBench e RULER mostram degradação de desempenho e maior sensibilidade à posição da evidência no prompt [Liu et al. 2024, Bai et al. 2025].

Nesse cenário, Sistemas Multiagentes (MAS) baseados em debate têm sido propostos como alternativa para melhorar raciocínio em tarefas complexas [Tang et al. 2025]. Contudo, ganhos de qualidade convivem com overhead de coordenação e desempenho inconsistente frente a baselines single-agent [Smit et al. 2024, Chen et al. 2024]. Assim, a pergunta central deixa de ser apenas *se* debate melhora a resposta, e passa a ser *quando* essa melhora compensa energeticamente.

Este trabalho torna explícita sua contribuição original em dois pontos. Primeiro, adotamos um critério metodológico rigoroso de equivalência qualitativa ($S_{\text{debate}} \geq S_{\text{single}}$): energia só é comparada quando o sistema multiagente preserva qualidade. Segundo, quantificamos empiricamente o *overhead extensional* com telemetria direta de hardware (NVML) em GPU H100, isto é, com medição observada em Joules e não por estimativas de software.

1.1. Trabalhos Relacionados (Síntese Analítica)

A literatura evoluiu em três frentes que raramente se cruzam. A primeira caracteriza a queda de qualidade em contexto longo [Liu et al. 2024, Bai et al. 2025]. A segunda mede custo energético de inferência em regimes longos [Wang et al. 2025, Wilkins et al. 2024]. A terceira investiga MAS/debate e aponta ganhos localizados, mas também overhead estrutural [Tang et al. 2025, Smit et al. 2024, Chen et al. 2024]. A lacuna está na interseção: poucos estudos integram qualidade e energia com telemetria direta sob critério explícito de equivalência qualitativa.

2. Design Experimental

2.1. Desenho Experimental

Comparamos duas arquiteturas sobre o mesmo conjunto de tarefas, mantendo modelo (Llama 3.2 2B) e hardware (NVIDIA H100) constantes: *Single Prompt* (monolítica) e *Debate Dialético* (multiagente). Foram executadas 10 rodadas completas por condição, totalizando 2.000 inferências.

2.2. Dados e Níveis de Complexidade

O dataset contém 100 tarefas do domínio financeiro (25 por nível), com contexto de 24k–32k tokens distribuído em quatro documentos corporativos (Cajueiro Bravo). A taxonomia de complexidade segue quatro níveis: N1 (recuperação factual), N2 (agregação), N3 (raciocínio multi-hop) e N4 (avaliação estratégica), alinhada ao framework *Depth/Width* [Bai et al. 2025].

2.3. Telemetria e Protocolo Estatístico

A energia foi medida por telemetria NVML (amostragem de 100 ms), por integração da potência no tempo. A qualidade (S) foi avaliada separadamente por veredito binário com LLM-judge (Mistral-3:14b), evitando contaminação da telemetria. Para energia, aplicamos Shapiro-Wilk e, diante de não normalidade, Mann-Whitney U com $\alpha = 0{,}05$, tamanho de efeito de Cliff's δ e correção de Holm-Bonferroni.

Os *system prompts*, o código experimental e a versão anonimizada do dataset estão disponíveis em repositório público para reprodutibilidade[1].

3. Resultados

3.1. Verificação de Equivalência de Qualidade

O pré-requisito metodológico para comparar energia é $S_{\text{debate}} \geq S_{\text{single}}$. A Tabela 1 resume os resultados por nível.

Tabela 1. Taxas de sucesso binário (S) por nível e elegibilidade para comparação energética.

Nível	$S_{\text{deb.}}$	$S_{\text{sing.}}$	Comp. energ.
N1	71,6% (179/250)	71,2% (178/250)	Sim
N2	1,2% (3/250)	0,0% (0/250)	Inconclusivo*
N3	2,0% (5/250)	5,6% (14/250)	Não
N4	2,0% (5/250)	7,6% (19/250)	Não

*N2: embora formalmente satisfaça $S_{\text{debate}} \geq S_{\text{single}}$, o resultado (3/250 vs. 0/250) é estatisticamente inconclusivo para comparações energéticas e não sustenta generalizações.

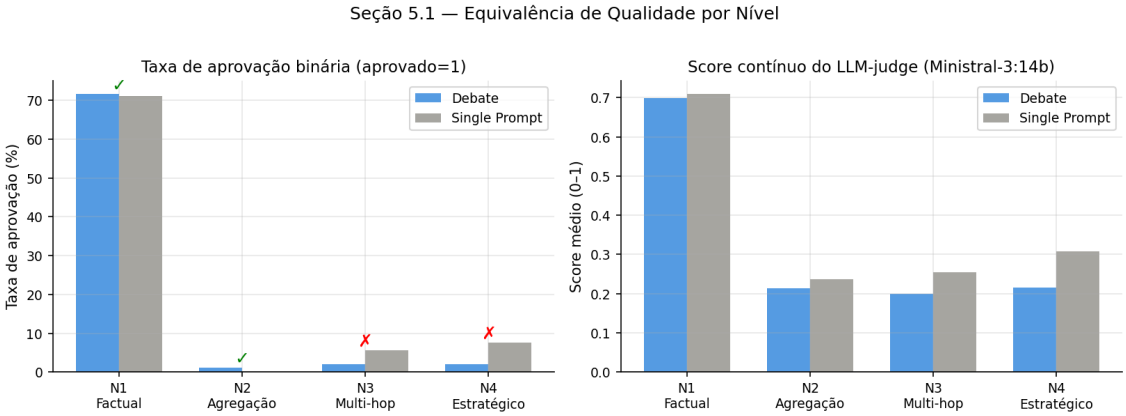

Figura 1. Equivalência de qualidade por nível, com taxa de aprovação binária e score contínuo do LLM-judge.

Tabela 2. Métricas energéticas globais (10 runs, 100 tarefas por arquitetura).

Métrica	Debate	Single	Razão (D/S)
Energia média por tarefa (J)	1069,6	136,5	7,84×
Mediana por tarefa (J)	1301,0	122,2	10,65×
Energia por token de saída (J/tok)	0,7668	0,9035	0,85×
Tokens de saída (média)	1394,9	151,1	9,23×
Latência média (ms)	9039	1127	8,02×
Potência média GPU (W)	118,3	120,7	0,98×

3.2. Comparação de Consumo Energético

O Debate apresenta maior consumo agregado (Mann-Whitney $p < 0{,}001$, Cliff's $\delta = 0{,}710$), com sobrecusto explicado principalmente pelo volume de saída e pela duração da execução.

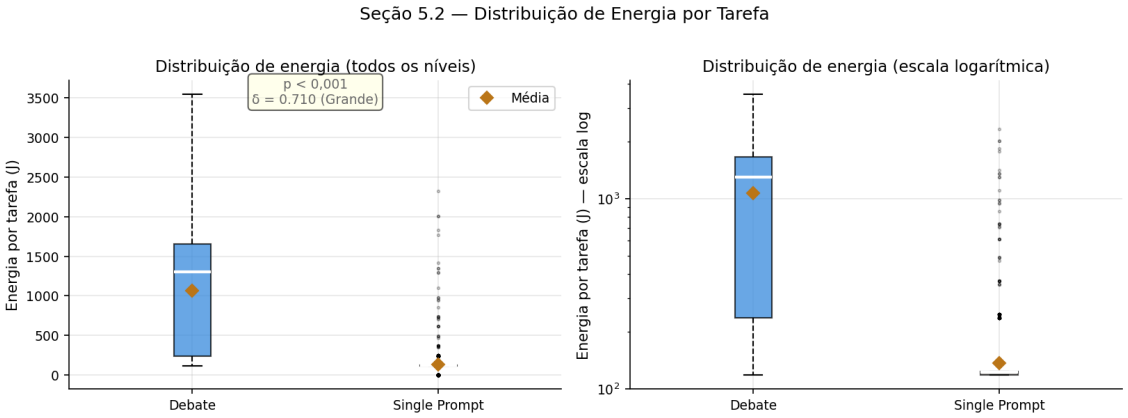

Figura 2. Distribuição de energia por tarefa (J), em escala linear e logarítmica.

3.3. Overhead Extensional e Limitações

O resultado central é um *overhead extensional*: a potência média da GPU é praticamente igual entre arquiteturas, mas o Debate executa por mais tempo devido ao volume de tokens gerados. Em N1, onde há equivalência qualitativa robusta, o Debate mantém APW apenas marginalmente superior; nos demais níveis, o custo adicional não se converte em ganho de qualidade.

Como esse overhead depende diretamente do volume de tokens, a tendência é que ele se mantenha (ou escale) em modelos maiores, agravando o custo energético absoluto do MAS.

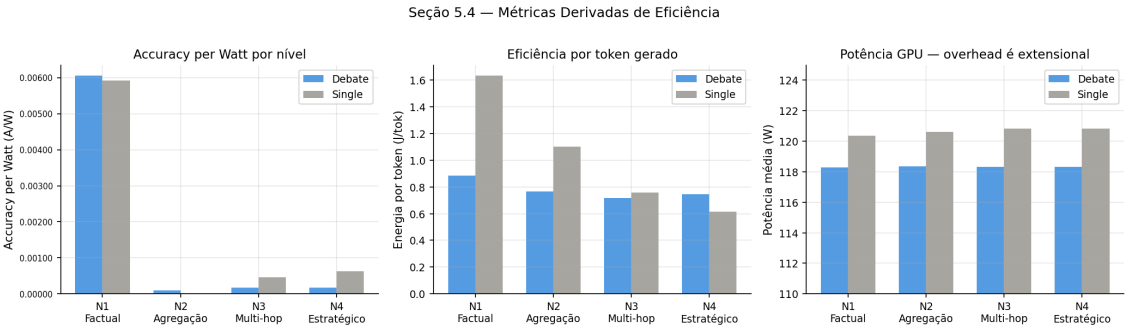

Figura 3. Métricas derivadas por nível; destaque para o painel de potência média da GPU (direita), que confirma regime de potência semelhante entre arquiteturas.

[1]Repositório: https://github.com/thiagoangelino/cajueiro-bravo-artifacts

3.4. Ameaças à Validade

As baixas taxas de sucesso em N3 e N4 (2,0% e 5,6%) podem refletir limitação cognitiva do Llama 3.2 2B para avaliação estratégica multi-hop, e não apenas efeito da arquitetura; futuros testes com modelos maiores (ex.: Llama 3 8B) são necessários para isolar esse fator. O uso do Mistral-3:14b como LLM-judge sobre o corpus sintético Cajueiro Bravo também pode introduzir viés sistemático na avaliação.

4. Discussão

4.1. Discussão Integrada

Os resultados indicam que o principal custo do Debate não é maior potência instantânea, mas maior duração de execução: a GPU opera em regime de potência semelhante, porém por mais tempo devido ao volume de tokens de coordenação. Esse padrão caracteriza um *overhead extensional*, consistente com a literatura sobre redundância de mensagens em MAS e com modelos token–energia da inferência.

Esse achado ajuda a explicar a divergência em relação à hipótese de ganho crescente com complexidade: nos níveis mais difíceis, o custo de coordenação cresce enquanto a qualidade não melhora de forma proporcional. Em termos práticos, a combinação de maior volume de saída e maior latência desloca o ponto de equilíbrio energético para tarefas mais simples.

Como o overhead observado é diretamente atrelado ao volume de tokens gerados, a tendência é que ele se mantenha (ou escale) em modelos maiores; nesse cenário, o custo energético absoluto do MAS tende a se agravar, mesmo quando a eficiência por token melhora marginalmente.

4.2. Implicações e Limitações

Para decisão de uso, os dados sugerem adotar Debate apenas quando houver evidência prévia de ganho qualitativo líquido no nível da tarefa. Sem esse ganho, o custo energético adicional tende a não se justificar. Como limitações, este estudo avalia uma configuração específica de Debate, um único modelo e um único hardware; estratégias adaptativas de poda/roteamento e execuções paralelas podem alterar a magnitude do overhead, mas não eliminam a necessidade de controlar volume de tokens como variável central. As baixas taxas de sucesso em N3 e N4 (2,0% e 5,6%) também podem refletir um gargalo cognitivo do Llama 3.2 2B para avaliação estratégica multi-hop, e não apenas efeito da arquitetura; trabalhos futuros devem incluir modelos maiores (por exemplo, Llama 3 8B) para isolar esse fator. Além disso, o uso do Mistral-3:14b como LLM-judge sobre um corpus sintético que ele não ajudou a gerar (Cajueiro Bravo) pode introduzir viés sistemático na avaliação.

5. Conclusão

Este trabalho avaliou empiricamente se arquiteturas de *Multi-Agent Debate* oferecem maior eficiência energética do que o *Single Prompt* em tarefas de raciocínio sobre contextos longos. A hipótese central — H_1: $E_{\text{multi}} < E_{\text{mono}}$ condicionado a $S_{\text{multi}} \geq S_{\text{mono}}$ — foi testada em 2.000 execuções com telemetria NVML em GPU H100 dedicada. H_1 **é rejeitada nas condições avaliadas.**

Nos níveis de maior complexidade cognitiva (N3 e N4), o pré-requisito de equivalência qualitativa é violado: o Single Prompt supera o Debate em acurácia binária ($5{,}6\%$ vs. $2{,}0\%$ em N3; $7{,}6\%$ vs. $2{,}0\%$ em N4). Nos níveis em que a equivalência é satisfeita (N1 e N2), o Debate consome entre $6{,}6\times$ e $7{,}0\times$ mais energia, com efeitos grandes e significativos (Cliff's $\delta \in [0{,}69,\ 0{,}74]$; $p < 10^{-38}$).

Três achados estruturam a contribuição do trabalho. Primeiro, o overhead é exclusivamente extensional: potências médias idênticas ($\Delta = -1{,}9\%$) confirmam que o custo adicional decorre de maior tempo de ocupação da GPU, não de maior demanda instantânea. Segundo, o volume de tokens de orquestração ($9{,}23\times$) é o principal *driver* do overhead, neutralizando a vantagem de $15{,}1\%$ em eficiência por token do Debate. Terceiro, a degradação qualitativa escala com a complexidade cognitiva — padrão oposto ao previsto por [Tang et al. 2025] e consistente com as críticas de [Smit et al. 2024] e [Chen et al. 2024].

A contribuição central é oferecer evidência empírica de hardware — não estimativas de software — de que a eficiência energética líquida de arquiteturas multiagentes depende criticamente do controle do volume de tokens de coordenação e da capacidade do modelo base de sustentar coerência em inferências distribuídas. Sob as condições avaliadas, o Single Prompt emerge como estratégia dominante em ambas as dimensões. A generalização desse resultado para arquiteturas com poda adaptativa de comunicação, modelos com maior capacidade de raciocínio distribuído e execução paralela de agentes constitui a agenda natural de trabalhos futuros.

Referências

Bai, Y. et al. (2025). LongBench pro: Towards harder tasks and more rigorous evaluation for long-context understanding.

Chen, L. et al. (2024). Are more LLM calls all you need? towards scaling laws of compound inference systems. In *NeurIPS 2024*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics*, volume 12, pages 157–173.

Smit, A. et al. (2024). Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *ICML 2024*.

Tang, X. et al. (2025). Multi-agent systems outperform single-agent on complex tasks: Scaling with reasoning depth.

Wang, H. et al. (2025). A systematic characterization of LLM inference on GPUs.

Wilkins, G., Keshav, S., and Mortier, R. (2024). Offline energy-optimal LLM serving: Workload-based energy models for LLM inference on heterogeneous systems. In *HotCarbon 2024*.