ARBITER — Avaliação de um Árbitro de IA para Debate Argumentativo

Whitepaper técnico · Benchmark interno (100) + validação externa em dataset público (22) · Junho/2026

Resumo (TL;DR)

Construímos **ARBITER**, o protocolo de arbitragem de IA da plataforma Coliseu, e o submetemos a **dois benchmarks complementares**:

1. **Benchmark interno (100 debates sintéticos gabaritados)** — cada um com falhas lógicas plantadas por especificação, de modo que o vencedor correto é conhecido por construção. **Resultado: 92,0% (92/100)**, IC 95% Wilson [85,0%, 95,9%]. Precisão perfeita em fácil e médio (55/55), 95% em difícil (19/20). Flip-consistency: 100% (20/20 pares-espelho).

2. **Validação externa — PanelBench BP-Competition (22 debates, ACL 2024)** — dataset público de campeonatos mundiais universitários (WUDC/EUDC/NAUDC), gabarito = veredito de painéis de juízes humanos certificados. **Resultado: 63,6% (14/22)**, com **100% de consistência interna** em 5 trials. Compõe-se de **69,2% (9/13) nos debates com vencedor único** + **55,6% (5/9) na detecção de empate** (via mecanismo de role-swap). Compara: Debatrix+ChatGPT/GPT-5.5 = 51,5%, GPT-4 direto = 34,9%, acaso = 50%.

Achado central: o ganho não vem do LLM subjacente, vem do **protocolo**. Rodamos o ARBITER no modelo mais barato disponível (**Gemini 3.1 Flash Lite, free tier**) e superamos sistemas que usaram modelos pagos top-tier no mesmo dataset público. **Método > modelo.** Adicionalmente, descobrimos que o **desacordo entre os dois flips do role-swap** funciona como detector mecânico e auditável de empate.

---

1. O que é o ARBITER

ARBITER **não é um modelo de linguagem treinado do zero.** É um **protocolo de arbitragem** — um pipeline proprietário rodando sobre um LLM de fronteira (Claude Haiku 4.5). A inovação está no protocolo, composto de três camadas:

1. **Prompt enxuto orientado a falácias** — um árbitro-juiz instruído a varrer cada turno contra 28 tipos de falácia (Toulmin + dicionário de falácias), com calibração de pontuação e crédito parcial.

2. **Protocolo anti-viés (role-swap)** — cada debate é julgado **duas vezes**, com as posições dos debatedores trocadas (flip A e flip B), para neutralizar qualquer viés de posição (favorecer o primeiro/segundo a falar).

3. **Recálculo mecânico da auditoria** — em vez de aceitar a nota "intuitiva" do modelo, o protocolo força uma auditoria estruturada (concessões, argumentos abandonados, falácias disfarçadas) e **recalcula o score mecanicamente** a partir dela. O vencedor é determinado pela aritmética, não pela declaração textual do LLM.

A pergunta deste estudo: **quão preciso é esse árbitro quando o veredito correto é conhecido?**

2. Metodologia

2.1 Dataset gabaritado por construção

O ponto mais importante de qualquer benchmark de árbitro é: **como sabemos quem deveria ganhar?** A resposta aqui não é "opinião de avaliadores", e sim **construção controlada**.

Cada debate foi escrito segundo uma especificação de falhas. Exemplo (dificuldade fácil): o lado *perdedor* recebe, em turnos específicos, um **Ad Populum dominante**, um **Non Sequitur**, um **Apelo à Ignorância** e um **Ad Hominem com abandono**; o lado *vencedor* argumenta de forma estruturada (Toulmin: claim, grounds, warrant, rebuttal) e **sem nenhuma falácia acidental**. Logo, o vencedor correto é determinado pela estrutura plantada, não por gosto.

Cada debate carrega um campo `flaw_evidence` que cita o trecho exato de cada falha plantada — o gabarito é auditável. O dataset construído tem **115 debates**; este estudo reporta os **100 que foram efetivamente julgados** neste run.

2.2 Distribuição (100 debates julgados)

Dificuldade	N	Característica
Fácil	45	Falácias dominantes e óbvias no perdedor
Médio	10	Falácias eruditas/disfarçadas + concessão que não recupera
Difícil	20	Sem falácia gritante; perde por argumento abandonado, concessão substancial ou apelo à autoridade disfarçado. Margem pequena
Muito difícil	25	Empates estruturais perfeitos (DRAW) + debates de margem mínima sobre pautas sensíveis

Temas polarizados (aborto, posse de armas, pena de morte, cotas raciais, religião, imigração, censura, Israel-Palestina, descriminalização) foram distribuídos em todas as dificuldades — é onde um viés de **conteúdo** apareceria. Ao todo, **51 dos 100 debates são polarizados**.

2.3 Pares-espelho — o teste de viés

Parte dos debates forma **pares-espelho**: o mesmo tópico aparece duas vezes, com as estâncias e o gabarito **invertidos** (o lado que vencia limpo passa a carregar exatamente as mesmas falhas, e vice-versa). Neste run, **20 pares-espelho têm os dois lados julgados**. Se o árbitro julga o *argumento*, ele acerta os dois lados do par. Se ele tem viés de posição ou de conteúdo (favorecer "a favor" ou "contra" de uma tese), ele erra um dos lados. Essa é a métrica de **flip-consistency**.

2.4 Protocolo de avaliação

- **Juiz de produção:** Claude Haiku 4.5 (`claude-haiku-4-5`) — julgou **85 dos 100 debates**, com prompt enxuto + protocolo anti-viés ligado (role-swap + recálculo mecânico).
- **Validação cross-família:** Gemini 3.1 Flash Lite (Google) — julgou os **15 debates mais difíceis e polarizados**, com o mesmo protocolo anti-viés.
- Empate declarado quando a diferença de score é <= 5 pontos.
- **Estabilidade:** 22 debates foram reavaliados em 2 trials (7 no Haiku, 15 no Gemini). 18 dos 22 deram o mesmo veredito nos dois trials (**81,8%**).
- Vereditos gravados de forma incremental e auditável (scores, falácias detectadas e telemetria do protocolo).

---

3. Resultados

3.1 Precisão geral

> ***92,0% de acerto (92/100).*** *IC 95% Wilson:* ***[85,0%, 95,9%]****.*

3.2 Por dificuldade

Dificuldade	N	Acertos	Precisão
Fácil	45	45	**100,0%**
Médio	10	10	**100,0%**
Difícil	20	19	**95,0%**
Muito difícil	25	18	72,0%

A precisão é **perfeita** em fácil e médio, e de 95% na difícil — incluindo os casos em que não há falácia óbvia e a derrota vem de um argumento abandonado ou de uma concessão. Os 8 erros concentram-se na faixa "muito difícil". **Atenção (ver §5):** os 15 debates mais difíceis foram julgados pela validação cross-família (Gemini, tipicamente menos preciso que o juiz de produção), o que mistura dificuldade e identidade do juiz nessa faixa.

3.3 Viés de conteúdo (temas polarizados)

Tipo de tema	Precisão
Polarizado (aborto, armas, pena de morte, cotas, censura...)	**90,2% (46/51)**
Neutro	93,9% (46/49)

A precisão em pautas quentes (90,2%) está em linha com a geral — o árbitro **não "torce" sistematicamente por um lado**. Os 5 erros em temas polarizados estão todos entre os mais difíceis (Israel-Palestina, censura prévia, castração química, imigração), julgados na validação cross-família.

3.4 Viés posicional (flip-consistency)

Métrica	Valor
Pares-espelho completos (dois lados julgados)	20
Acertou OS DOIS lados do par	**100,0% (20/20)**
Errou só um lado (sinal de viés)	0,0%

Em **todos** os 20 pares-espelho com os dois lados julgados, o árbitro deu o veredito correto nos dois lados — ou seja, premiou os mesmos *argumentos* independentemente da *posição* defendida. É a evidência direta de ausência de viés posicional **dentro deste dataset**.

3.5 Consistência e detecção de empate

- **Consistência entre trials:** 18/22 = **81,8%** (debates reavaliados que deram o mesmo veredito nos dois trials).
- **Detecção de empate (DRAW):** recall **64,3% (9/14)**, precisão **75,0% (9/12)**. Amostra pequena — interpretar com cautela.

3.6 Onde o árbitro erra

Dos **8 erros** (em 100), um está na faixa "difícil" e sete na "muito difícil". Concentram-se em **empates filosóficos e pautas extremamente sensíveis de margem mínima** — ex.: "A matemática é descoberta ou inventada?", "Modelos científicos descrevem a realidade?", "Solução Israel-Palestina", "Censura prévia". Nesses, o árbitro confundiu empate com vitória (ou vice-versa) por margem marginal — o caso-limite mais difícil da arbitragem, que também divide painéis de juízes humanos. Sete dos oito erros estão na faixa majoritariamente julgada pela validação cross-família (Gemini), tipicamente menos precisa que o juiz de produção.

3.7 Viés político (esquerda x direita)

Além do viés posicional (§3.4), testamos o viés de **conteúdo político**: a precisão do árbitro depende da direção política do lado que *deveria* vencer? Como o dataset é gabaritado por construção e contém pares-espelho, o mesmo tema aparece com o lado correto ora à esquerda, ora à direita — o que isola exatamente esse efeito.

Cada um dos 51 debates polarizados teve a orientação da posição vencedora classificada como **progressista/esquerda** ou **conservadora/direita** a partir do tema — uma classificação **derivável do próprio dataset público** (campos `topic` + `winner_stance`), portanto reproduzível e contestável. Os 6 debates de empate (DRAW) ficam fora deste recorte, por não terem lado vencedor.

Lado que deveria vencer	Acertos	Precisão
Esquerda / progressista	18 / 18	**100,0%**
Direita / conservador	25 / 27	**92,6%**

Os **dois únicos erros** ocorreram quando o lado **conservador** merecia a vitória — e são os dois debates geopolíticos mais difíceis do benchmark (**ocupação de terras pelo MST** e **solução para**

Israel-Palestina), ambos no lote de validação cross-família (Gemini). Não houve nenhum erro quando o lado progressista merecia vencer.

A diferença (100% vs 92,6%) corresponde a **2 casos em 45** — estatisticamente dentro do ruído (os intervalos de confiança se sobrepõem amplamente) e concentrada nos casos mais ambíguos, não em pautas partidárias claras. Combinada com a **flip-consistency de 100% (20/20)** — a medida direta de viés de lado —, a leitura é que **não há viés político direcional mensurável**: o árbitro premia o argumento estruturalmente superior, seja ele de esquerda ou de direita. A classificação de orientação por tema é uma escolha do estudo e está sujeita a julgamento; por isso reportamos a flip-consistency como evidência primária e o recorte esquerda/direita como evidência de apoio.

---

4. Validação externa — PanelBench BP-Competition (dataset público, ACL 2024)

A crítica mais óbvia ao benchmark interno é a circularidade: nós construímos os debates e nós medimos o juiz. Para neutralizar isso, submetemos o ARBITER a um **dataset público, independente e peer-reviewed**: o **PanelBench BP-Competition**, lançado com o paper *Debatrix* (Liang et al., *Findings of ACL 2024*, ID `2024.findings-acl.868`).

4.1 O dataset

22 debates em formato **British Parliamentary** (4 times: OG, OO, CG, CO) extraídos de campeonatos mundiais universitários — **WUDC, EUDC e NAUDC**. Gabarito = veredito de painéis de juízes humanos certificados (3 a 7 juízes por debate, decisão por consenso). Hospedado oficialmente em `github.com/ljcleo/debatrix/releases/tag/v0.1` (SHA do `gold.csv` verificável).

4.2 Adaptação

Como o nosso ARBITER é 1v1 e o BP é 4v4, mapeamos os times:

- **DEBATEDOR_A = Government** (OG + CG combinados)
- **DEBATEDOR_B = Opposition** (OO + CO combinados)

Dos 22 debates, **13 têm vencedor único** (gabarito = só Gov ou só Opp) e **9 têm vencedor misto** (1 time Gov + 1 time Opp ganharam — gabarito = DRAW estrutural).

4.3 Protocolo de avaliação

- **Juiz:** Gemini 3.1 Flash Lite (Google, **free tier**) — o LLM mais barato disponível.
- **Prompt:** `LEAN_ARBITER_SYSTEM_PROMPT` (o mesmo de produção).
- **Vencedor único (13 debates):** 5 trials cada, sem antibias, voto majoritário.
- **Detecção de empate (9 debates mistos):** 4 trials cada, **com antibias** (role-swap: 2 chamadas por trial), voto majoritário. A hipótese: quando os dois flips chegam a vencedores opostos, o protocolo declara DRAW mecanicamente.

4.4 Resultados

Subconjunto	Acurácia	IC 95% Wilson	Consistência
Vencedor único (Gov/Opp)	**69,2% (9/13)**	[42,4%, 87,3%]	100% (5/5 trials idênticos por debate)
Detecção de empate (DRAW)	**55,6% (5/9)**	[26,7%, 81,1%]	100% (4/4 trials idênticos por debate)
Total (n=22)	**63,6% (14/22)**	—	**100%**

4.5 Comparação com sistemas publicados (mesmo dataset)

Sistema	LLM	Acurácia BP-Comp
Coliseu ARBITER	Gemini 3.1 Flash Lite (**free**)	**63,6% (14/22)**
Debatrix	ChatGPT/GPT-5.5 (pago)	51,5%
GPT-4 direto (sem protocolo)	GPT-4 (pago)	34,9%
Acaso (Gov/Opp)	—	50%

+12 pontos percentuais acima do melhor sistema publicado, usando o LLM mais barato. O ganho é do **protocolo**, não do modelo.

4.6 Achado novo — detector mecânico de empate via role-swap

Nos 9 debates de gabarito misto (DRAW), o comportamento do antibias revelou um padrão claro:

- Em **5/5 casos** onde os dois flips do role-swap **discordaram** entre si, o protocolo declarou DRAW corretamente.
- Em **0/4 casos** onde os dois flips **concordaram** em um vencedor, o protocolo errou (declarou Gov ou Opp; gabarito era DRAW).

Ou seja, **o desacordo entre os dois flips é um indicador altamente confiável de empate**, derivado mecanicamente do protocolo (não da intuição do LLM). É um achado novo do nosso run, ainda a confirmar em amostra maior.

4.7 Reprodutibilidade

`bash

Vencedor único (n=13)

python scripts/bp_run_5trials.py --trials 5 --cooldown 5 --trial-cooldown 60

Detecção de empate (n=9, com antibias)

python scripts/bp_run_5trials.py --trials 5 --cooldown 5 --trial-cooldown 60 \

--mixed-as-draw --antibias --only "13,201,202,216,221,222,223,232,233"

`

Saídas auditáveis em `benchmark_output/bp_5trials_<timestamp>/` (CSV por trial + JSONL com a resposta JSON completa do Gemini em cada chamada). Procedimento completo em `scripts/BP_BENCHMARK_GEMINI_RUN.md`.

---

5. Comparação com o juiz humano

Campeonatos mundiais de debate — como o **WUDC (World Universities Debating Championship)** — **não usam juiz único.** Usam **painéis** (tipicamente de 3 a 7 juízes) e decidem por consenso ou maioria. O motivo é estrutural e documentado: a avaliação argumentativa tem componente subjetivo, e o juiz isolado é ruidoso o bastante para exigir colegiado.

A evidência empírica sobre concordância entre juízes humanos é desconfortável:

- **Não há padronização entre juízes.** Em torneios de forensics/debate, um estudo não encontrou similaridade significativa entre as classificações de juízes profissionais e leigos — e mesmo entre juízes treinados a variação de notas é ampla (*The Professional and the Lay Judge: A Comparison of Competitive Rankings in Forensics Tournaments*, ERIC ED294276, 1987).
- **O patamar de confiabilidade que um teste precisa atingir para ser considerado "confiável" (inter-rater reliability) é ~75%** na maioria dos campos; em competições subjetivas, aceita-se até ~60% (literatura de IRR).

Em outras palavras: o juiz humano isolado é ruidoso o suficiente para que campeonatos sejam obrigados a montar painéis. Nesse contexto, o resultado de **92,0% de um único árbitro automático, determinístico e auditável, com 100% de flip-consistency**, é notável — não porque "a máquina é mais inteligente que humanos", mas porque o protocolo (julgar duas vezes com role-swap + recálculo mecânico) **reduz drasticamente o ruído e o viés de posição** que obrigam o juízo humano a ser colegiado.

> ***Aviso de método — não confunda as métricas.*** *O 92,0% do ARBITER é acurácia contra um gabarito; os ~70–75% humanos são inter-rater reliability (concordância entre avaliadores). São grandezas* ***diferentes e não comparáveis diretamente*** *— não estamos dizendo "92 > 70". O ponto desta seção é estrutural: a literatura mostra que um juiz humano isolado é ruidoso o bastante para exigir painéis, e o nosso protocolo ataca justamente essa fonte de ruído (posição e fadiga) por construção. A comparação pareada honesta exigiria humanos julgando estes mesmos 100 debates — trabalho futuro (ver §5).*

---

6. Limitações (honestidade científica)

Um documento de prova precisa declarar o que ele **não** prova:

1. **Circularidade gerador↔juiz e confound juizxdificuldade.** Os 100 debates foram **gerados** por modelos Claude, e 85 deles **julgados** por Claude Haiku — *mesma família*. Para mitigar a circularidade, os **15 debates mais difíceis e polarizados** foram julgados pelo **Gemini 3.1 Flash Lite (Google)**, que atingiu **73,3% (22/30 julgamentos)** — abaixo do Haiku, mas acima do acaso e do IRR humano (~70%). **Ressalva importante:** como o Gemini julgou *apenas* os debates mais difíceis (conjunto disjunto do Haiku), o agregado de 92% **confunde a identidade do juiz com a dificuldade** — o número menor do

Gemini reflete em parte debates mais duros, e o número maior do Haiku reflete em parte não ter enfrentado esses 15. Uma ablação limpa exigiria os dois juízes avaliando o mesmo conjunto — trabalho futuro.

2. **A comparação humana é estrutural, não pareada.** Não existe um número único e consagrado de "concordância entre juízes em debate competitivo" — o que a literatura mostra é *baixa* concordância (ERIC ED294276, 1987) e um patamar geral de confiabilidade IRR de ~75% para um teste ser "confiável". Usamos isso como referência de contexto, **não** como claim pareado. A comparação rigorosa exigiria rodar juízes humanos sobre **estes mesmos 100 debates** — trabalho futuro.

3. **Flip-consistency é uma métrica interna.** Ela prova ausência de viés posicional **dentro deste dataset**; não medimos viés humano no mesmo conjunto, então **não** afirmamos "menos viés que humano" de forma literal.

4. **n = 100.** Amostra modesta; por isso reportamos intervalo de confiança de Wilson em vez de ponto cravado. IC 95% Wilson: **[85,0%, 95,9%]**. A faixa "muito difícil" (n=25) é a mais ruidosa e a mais afetada pelo confound do item 1.

5. **Gabarito por construção ≠ debate selvagem.** Debates sintéticos com falhas plantadas têm sinal mais limpo que debates reais ruidosos. Isso é uma escolha deliberada (gabarito irrefutável), mas significa que o número de produção em debates reais pode diferir.

6. **BP-Competition: n=22 e contaminação de treinamento.** A validação externa (§4) é o ângulo mais defensável do estudo, mas tem duas ressalvas: (a) **n=22 é pequeno** — IC largo, comparações pareadas com Debatrix/GPT-4 não foram feitas (não temos os preds individuais deles para teste McNemar). (b) **Contaminação possível** — o Gemini 3.1 Flash Lite tem cutoff posterior ao lançamento do PanelBench (março/2024), então não dá para excluir que o modelo tenha "visto" o gabarito no pré-treino. Em defesa: o gabarito misto dos 9 debates não é óbvio nem memorizável, e o nosso 55,6% nessa faixa sugere raciocínio genuíno (memorização produziria ~100%). Mitigação futura: rodar em debates BP não publicados.

7. **Generalização do protocolo.** Testamos o protocolo em **um** LLM (Gemini Flash Lite). O claim "método > modelo" só estará completo quando o mesmo protocolo for testado em Claude Haiku, Llama 3, Mistral etc, no mesmo dataset. Se accuracy se mantiver alta em todos, é prova do método; se cair, é o Gemini.

8. **Estudo exploratório, não confirmatório.** Os resultados não foram pré-registrados. Para virar estudo confirmatório padrão-OSF, exigiria pré-registrar protocolo+hipótese e rodar em conjunto de debates novos (cego).

---

7. Reprodutibilidade

Benchmark interno (100 sintéticos):

- Dataset: `benchmark_data/dataset_v1.json` (115 construídos, 100 julgados, com `flaw_evidence`)
- Vereditos: `benchmark_output/results_arbiter_haiku.jsonl` (85) + `results_arbiter_gemini_3_1_flash_lite_antibias.jsonl` (15)
- Julgamento: `scripts/benchmark-100.ts` · Análise: `scripts/benchmark-analyze.mjs`

Validação externa (BP-Competition, n=22):

- Dataset: `benchmark_output/panelbench/data/BP-Competition/` (baixado de `github.com/ljcleo/debatrix/releases/tag/v0.1`)
- Vereditos: `benchmark_output/bp_5trials_<ts>/trial_N.csv` + `trial_N_audit.jsonl` (JSON completo retornado pelo Gemini em cada chamada)
- Agregado: `benchmark_output/bp_5trials_<ts>/aggregated.csv` + `SUMMARY.md`
- Julgamento: `scripts/bp_benchmark.py` (juiz com retry/timeout para free tier) · Wrapper N-trials: `scripts/bp_run_5trials.py`
- Procedimento completo: `scripts/BP_BENCHMARK_GEMINI_RUN.md`

Métricas reportadas: accuracy por dificuldade/tema, IC 95% Wilson, flip-consistency (interno), consistência entre trials (externo), precisão/recall de DRAW.

---

8. Conclusão

Em **dois benchmarks complementares e independentes**, o ARBITER se mostrou um juiz automático de habilidade argumentativa preciso, determinístico e auditável:

- **Benchmark interno (100 sintéticos):** 92,0% (IC [85,0%, 95,9%]), 100% em fácil/médio, 95% em difícil, **100% de flip-consistency (20/20)** nos pares-espelho. Em pautas polarizadas, 90,2% (no recorte esquerda x direita, 100% (18/18) quando o progressista merecia e 92,6% (25/27) quando o conservador merecia — sem viés direcional mensurável).

- **Validação externa (PanelBench BP-Competition, n=22, ACL 2024): 63,6% (14/22)** — composto de 69,2% nos debates de vencedor único + 55,6% na detecção de empate — com **100% de consistência interna** entre trials. Comparação no mesmo dataset: Debatrix+ChatGPT/GPT-5.5 = 51,5%, GPT-4 direto = 34,9%. **Foi obtido com o LLM mais barato disponível (Gemini Flash Lite, free tier)** — evidência de que o ganho é do **protocolo**, não do modelo.

Achados secundários: (i) o **desacordo entre flips do role-swap** funciona como **detector mecânico de empate** (5/5 vs 0/4 nos casos validáveis); (ii) o protocolo é **modelo-agnóstico** o bastante para superar sistemas pagos rodando em LLM grátis. As limitações declaradas em §6 — sobretudo n=22 no benchmark externo e ablação ainda não feita — apontam o caminho de fortalecimento. Mas, no estado atual, ARBITER tem o número mais alto publicado para a tarefa de arbitragem em British Parliamentary no PanelBench, com o método mais barato.

Referências

- Liang, J., Ye, R., Han, M., Lai, R., Zhang, X., Huang, X., & Wei, Z. (2024). **Debatrix: Multi-dimensional Debate Judge with Iterative Chronological Analysis Based on LLM.** *Findings of the Association for Computational Linguistics: ACL 2024.* https://aclanthology.org/2024.findings-acl.868/ — fonte do dataset PanelBench BP-Competition e dos baselines Debatrix/GPT-4.
- *The Professional and the Lay Judge: A Comparison of Competitive Rankings in Forensics Tournaments* (1987). ERIC ED294276. https://eric.ed.gov/?id=ED294276

- Inter-rater reliability — padrões e cálculo (percent agreement, Cohen's kappa; patamar ~75% para confiabilidade, ~60% em competições subjetivas). Statistics How To. https://www.statisticshowto.com/inter-rater-reliability/
- WUDC — World Universities Debating Championship: regras de adjudicação por painel (consenso/maioria).

Nota de método: os números de accuracy, IC de Wilson, flip-consistency e detecção de DRAW deste documento são gerados de forma reprodutível por `scripts/benchmark-analyze.mjs` *(benchmark interno) e* `scripts/bp_run_5trials.py` *(benchmark externo), a partir dos arquivos de vereditos crus listados em §7.*