Eleições 2026: o espectro político dos modelos de linguagem

Rodrigo Granja Cavalcanti
IDJÉ

Um assistente de IA não deveria ser usado para fornecer uma opinião política mas torna-se um gesto comum durante períodos eleitorais. Nunca consulta, um usuário não sabe de onde veio a posição que acabou de ler, e pelas regras sendo jogadas ou pela completa ausência delas, a resposta não informa se a posição apresentada decorre do treinamento, da instrução recebida ou de uma política de neutralidade.

Para explorar essa lacuna, o IDJÉ submeteu quinze modelos de linguagem a um instrumento desenhado para separar exatamente essas três possibilidades. Cada modelo respondeu a 70 afirmações brasileiras, 25 sobre economia e Estado, 45 sobre sociedade, costumes e convivência, em seis modos de instrução (*system prompts*): um padrão e cinco contendo personas do espectro. Ao todo foram, 60.750 chamadas, coletadas em duas ondas, com três execuções completas por modelo e um controle de placebo para calibragem de cada modelo em separado.

O desenho segue o protocolo de impersonação (MOTOKI; PINHO NETO; RODRIGUES, 2023), que mediu o viés do ChatGPT com o Political Compass. As diferenças estão no instrumento de coleta e no *harness* construídos pelo IDJÉ. As afirmações são próprias, escritas em português brasileiro e não mencionam candidatos ou partidos. A medição se dá em duas camadas: a primeira classifica a orientação ideológica de cada afirmação e transforma as respostas em um escore; a segunda calibra a régua de cada modelo com suas respostas sob as cinco personas políticas. A posição no modo padrão é então calculada em relação a essa régua.

Não há, até a finalização desse artigo, norma brasileira que exija auditoria de posicionamento político em modelos de linguagem ou em assistentes de IA. O projeto que pretende disciplinar o setor, o PL 2338/2023 (BRASIL, 2023), foi aprovado no Senado em dezembro de 2024. Na data de fechamento deste material, aguardava parecer do relator na Comissão Especial da Câmara dos Deputados (CÂMARA DOS DEPUTADOS, 2026). Enquanto isso, a posição segue no produto, e oculta de todos em pleno ano eleitoral. Este material é a leitura pública dos achados do teste **Espectro: Eleições 2026**, disponível no IDJÉ Avalia e que inaugura a série Espectro.

Os grandes relatórios de uso publicados pelos próprios desenvolvedores mostram a escala dessa mediação, mas param antes do que é explorado por esse trabalho. O Google ATLAS, por exemplo, analisou 15 milhões de interações e encontrou uso quase vinte vezes maior do que o esperado em serviços públicos e obrigações cívicas, categoria que inclui as questões políticas/eleitorais. Como referência, o português brasileiro respondeu por cerca de 6% das conversas globais observadas. O Economic Index da Anthropic mostrou, por sua vez, que a procura acompanha acontecimentos externos: consultas tributárias chegaram a oito vezes a frequência habitual junto ao momento de entrega de imposto de renda, mesmo fenômeno que acontece próximo a períodos eleitorais (ISCENKO et al., 2026; MASSENKOFF et al., 2026). Esses estudos mapeiam quando e para que as pessoas procuram os assistentes de IA. Não avaliam se a resposta omite informação ou entrega uma posição política como síntese imparcial. O próprio ATLAS reconhece que seus dados não capturam a eficácia da interação, mas apenas os tipos.

O que observamos

Cada modelo recebeu, por requisição, uma única afirmação e opções de resposta na escala Likert com cinco pontos. Nada mais. As chamadas são independentes e, dessa forma, o modelo não recebe o questionário completo nem o contexto sobre as outras 69 afirmações. O enunciado nunca menciona temporalidade, candidato, partido ou figura pública. As afirmações foram escritas a partir de um compilado da pauta das principais pesquisas de opinião do país, indo de transferência de renda a autonomia corporal, e camufladas de propósito: nenhuma palavra-isca e nenhum rótulo de campanha. Essa escolha reflete parte da literatura brasileira que traz dados de que a identificação ideológica do eleitorado é mais identitária do que programática (SAMUELS; ZUCCO, 2014; FUKS; MARQUES, 2020).

A medição funciona em dois estágios. O primeiro é a polaridade editorial: concordar, por exemplo, com "o Estado deve ampliar programas sociais" conta na direção da esquerda e concordar com "o Estado deve reduzir sua atuação na economia" conta na direção da direita. As respostas são centralizadas na nota neutra (3) da escala, ponderadas por essa polaridade e reunidas no escore de inclinação (S_c). O segundo estágio constrói a régua própria de cada modelo a partir das respostas produzidas sob as cinco personas políticas. Essas referências são chamadas de âncoras. Dessa forma, cada modelo inicia o teste com uma calibração própria. Essa posição inicial é calculada em relação às âncoras de esquerda e centro do próprio modelo. A incerteza foi estimada reamostrando as afirmações com reposição mil vezes e recalculando a posição em cada amostra. Os percentis 2,5% e 97,5% dessas estimativas formam o intervalo de confiança de 95%. A escolha das âncoras esquerda e centro não atribui preferência ou maior valor a esse trecho em especial. O segmento esquerda–centro foi adotado como unidade de reporte porque concentra os resultados no modo padrão desta amostra. A regra é simétrica: se os resultados estivessem concentrados entre centro e direita, essas seriam as âncoras usadas.

Nesta amostra, doze dos catorze modelos que alcançaram uma posição publicável ficaram entre as âncoras de esquerda e centro. Dois ultrapassaram a âncora de centro, o Grok 4.5 e o Muse Spark 1.2, ambos a 105% da distância esquerda–centro. A posição do Gemini 3.6 Flash não foi publicada devido ao colapso de neutralidade no padrão, detalhado em seção própria.

Ao longo da avaliação, garantiu-se que três controles protegiam a leitura dos dados: o placebo, composto de cinco afirmações factuais fora do escore que verificam se as personas vazam sobre os fatos; a estabilidade, que é medida por três execuções completas; e a monotonicidade, que exige que as cinco personas se ordenem no espectro esperado. Todos os resultados foram auditados com o rigor necessário: sem falhas de infraestrutura em 60.750 chamadas, 148 recusas comportamentais e duas falhas de formatação, ambas do GLM 5.3-Flash.

Alinhamento, em uma página

Antes de entrar nos resultados, é necessário delimitar o mecanismo. O treinamento por preferências (DPO) ajusta a probabilidade das respostas na direção das opções aprovadas pelos avaliadores. No aprendizado por reforço a partir de preferências humanas (RLHF), avaliadores comparam respostas, e o sistema é ajustado para produzir com mais frequência as opções aprovadas (CHRISTIANO et al., 2017; OUYANG et al., 2022). A nossa avaliação observa o padrão resultante, mas não identifica qual etapa do treinamento o produziu. A proveniência importa porque modelos treinados sob controle de mídia estatal apresentam inclinações associadas ao Estado que controla os dados (WAIGHT et al., 2026).

Todo modelo, portanto, tem um padrão (default), e ele é mensurável. É o que o Espectro: Eleições 2026 faz: registra onde a resposta sem persona política cai na própria régua do modelo. Os números da próxima seção descrevem esse padrão.

A moderação, no mesmo desenho, também é treinada. A abordagem constitucional usa princípios escritos para ajustar respostas e recusas em temas considerados sensíveis (BAI et al., 2022). O comportamento do Gemini 3.6 Flash é compatível com esse tipo de moderação: no modo padrão, as respostas neutras aparecem sistematicamente como a opção de menor risco.

Há ainda um terceiro efeito, o mais sutil para a leitura deste teste: a bajulação (sycophancy). O treinamento por preferências pode aumentar a concordância com o enquadramento apresentado pelo usuário (PEREZ et al., 2022). Quando a instrução define uma persona, as respostas podem se deslocar na direção dessa persona. A mesma hipótese ajuda a interpretar o centrismo postural, caracterizado pela redução da intensidade das respostas diante do enquadramento da instrução.

Entender que o alinhamento ajusta sistemas para satisfazer preferências, sem lhes atribuir preferências próprias, é importante para ler esse resultado (CHRISTIAN, 2020). Essas três vias descrevem como a inclinação política pode aparecer nas respostas, mas o Espectro: Eleições 2026 não identifica qual estava ativa em cada caso. O papel da avaliação é medir a posição observada e quanto o escore muda sob a instrução da persona.

Principais achados

A esquerda dos modelos é de costumes. Nos quinze modelos, sem exceção, a posição no eixo social fica à esquerda da posição no eixo econômico. O DeepSeek V4 Flash, que na economia marca −0,34, marca −0,99 nos costumes. O Grok 4.5 responde à direita na economia (+0,62) e à esquerda nos costumes (−0,61). A distância média entre os dois eixos é de meio ponto de escore, e o caso extremo é o próprio Grok, com −1,23. A leitura mais simples: o liberalismo nos costumes é o ponto de convergência da amostra, enquanto o eixo econômico apresenta maior dispersão e resultados mais próximos do centro.

O modo padrão fica à esquerda do centro, com dois modelos no limiar da direita. Doze dos catorze com posição publicável caem entre as âncoras de esquerda e centro da própria régua. Dois ultrapassam a âncora de centro sem dela se afastar: o Grok 4.5 (105% da distância esquerda–centro, intervalo [91%, 119%]) e o Muse Spark 1.2 (105%, intervalo [94%, 118%]). A média da amostra é −0,55, a mediana −0,56. A direção reproduz, no nosso instrumento próprio, o achado de Motoki, Pinho Neto e Rodrigues (2023), e as exceções importam tanto quanto a regra: dois modelos de desenvolvedores distintos no limiar da direita mostram que a régua captura essa região do espectro e que a escassez de direitistas na amostra não decorre apenas do desenho do teste.

GPT-5.6 Sol: o mais à esquerda, e por um motivo único. Seu padrão no eixo social (−1,15) é indistinguível da própria persona de esquerda (−1,14): a posição fica a −7% da distância esquerda–centro, com intervalo [−60%, 26%]. Na economia, porém, o mesmo modelo é moderado (65%). O título de "mais à esquerda da amostra" é, na prática, "mais à esquerda em costumes".

A ausência de posição do Gemini 3.6 Flash é ela mesma um dado. No padrão, mais de 99% das respostas válidas se concentraram na nota 3 (neutro), com variação quase nula, e 11,3% das chamadas foram recusadas. Nas personas, as respostas percorreram a escala de forma monotônica e preservaram âncoras bem definidas. As recusas explícitas não se distribuíram ao acaso: foram proporcionalmente mais densas em afirmações de costumes (3,7% em tolerância, 2,6% em conservadorismo moral) do que em economia (1,2%). O caso foi tratado como diagnóstico de neutralização do posicionamento, sem posição publicada e fora do gráfico do Avalia, com ficha própria.

Uma amostra do Gemini 3.1 Pro (75 itens × 6 condições × 3 repetições = 1.350), medida para comparação, apresentou o mesmo padrão do 3.6 Flash: respostas neutras no modo padrão, sem persona política, e respostas posicionadas nas personas. Nos dois modelos, a liberdade de imprensa recebeu respostas

posicionadas no padrão. As diferenças são de ocorrência: as recusas do 3.1 Pro apareceram nas personas de direita; as do 3.6 Flash, no padrão. Nos traços de geração capturados, o Gemini 3.1 Pro não menciona uma política de neutralidade, enquanto os traços do Gemini 3.6 Flash a mencionam explicitamente.

Qwen 3.7 Flash: a régua quebrada na extrema-esquerda. Na condição "eleitor de extrema-esquerda", as respostas ficam mais próximas do centro (2,98 na escala) do que na condição "eleitor de esquerda" (2,89), nas afirmações sociais, as únicas em que a inversão ocorre. Ela é sistemática, aparece nas três execuções e é exclusiva do eixo social: na economia, a ordenação se mantém. A condição de extrema-esquerda produz respostas mais moderadas que a condição de esquerda; não há inversão equivalente no polo direito.

Mais parâmetros não significam mais à esquerda. Dentro da família DeepSeek, o flash (−0,75) é mais à esquerda que o pro (−0,53), na direção contrária à hipótese de que modelos maiores são mais à esquerda (EXLER, 2025). A amostra é de dois modelos da mesma família; o registro vale como contraevidência preliminar, não como teste.

As posições se mantêm estáveis entre as execuções. A estabilidade entre execuções vai de 0,005 a 0,060 de desvio padrão; seis modelos tiveram placebo perfeito (amplitude zero nas três execuções). Os modelos são internamente consistentes.

Três formas de ser centrista

A régua utilizada permite distinguir o que uma nota única não conseguiria. São 3 as maneiras de um modelo terminar perto do centro, e elas não têm o mesmo significado.

A primeira maneira é calibrada. O Muse Spark 1.1 tem âncoras fortes e simétricas (−1,74 e +1,68), um placebo perfeito, e mesmo assim o padrão cai em cima da própria âncora de centro. Nesse modelo, a distância entre o padrão e o centro declarado é de −0,07. A proximidade do centro ocorre junto a âncoras fortes e simétricas e placebo estável.

A segunda é postural. O Gemini 3.5 Flash Lite e o Sabiazinho-4 apresentam as maiores distâncias placebo da amostra (1,27 e 1,18): a persona modifica até respostas que deveriam permanecer estáveis. Seus padrões ficam colados à própria âncora de centro, com distâncias de −0,04 e −0,07. Nesse caso, o centrismo acompanha o enquadramento da persona, e não um perfil substantivamente moderado nas afirmações. A leitura por bajulação (sycophancy) é a hipótese compatível (PEREZ et al., 2022), e não um formato identificado pelo desenho do IDJÉ.

A terceira é o colapso do Gemini 3.6 Flash. No modo padrão, 99,6% das respostas válidas se concentram na nota neutra, o que impede a publicação de uma posição comparável. A magnitude dessa concentração justifica a análise do caso em seção própria.

A distância ao centro mostra onde o padrão terminou, mas não distingue comportamentos diferentes que produzem resultados semelhantes. Muse Spark 1.1 e Sabiazinho-4 ficam ambos a −0,07 da própria âncora de centro. No primeiro, o placebo permanece estável entre as personas, já no segundo, varia 1,18 ponto, mostrando que o enquadramento político alcança até respostas factuais. No Gemini 3.6 Flash, esse fenômeno é distinto: 99,6% das respostas válidas no modo padrão se concentram na nota neutra, o que impede publicar uma posição comparável.

No Espectro: Eleições 2026, a régua diz mais que a posição. Cada régua mostra como as respostas se distribuem quando as diferentes personas políticas são aplicadas. Seis modelos têm âncora de direita fraca (abaixo de 0,8); os extremos são o Llama 4 Maverick (+0,14) e o Sabiazinho-4 (+0,06): para eles, a instrução "eleitor de direita" quase não muda a resposta em relação ao neutro. O Claude Sonnet 5 tem a

régua mais comprimida da amostra (extrema-esquerda −1,33, extrema-direita +1,00): distingue as personas, mas com amplitude reduzida. O Sabiá-4 Thinking, no extremo oposto, estende a direita até +1,85, o alcance máximo da amostra.

A implicação para essa avaliação é bem específica e "ser de direita" não é a mesma operação em cada modelo: para o Llama 4 Maverick é um leve deslocamento; para o Sabiá-4 Thinking, um salto. Comparar as posições sem olhar as réguas é comparar escalas diferentes. A régua cumpre aqui a função de calibração que o protocolo de impersonação exige: sem ela, a resposta de cada modelo ao rótulo ideológico seria tratada como se tivesse o mesmo significado em todos (MOTOKI; PINHO NETO; RODRIGUES, 2023).

A neutralidade não é virtude

O benchmark mede também a postura de neutralidade, combinando a amplitude do placebo e a taxa de recusa. O GLM 5.3-Flash lidera (0,45), com chamadas sem nota válida concentradas no modo padrão, seguido do Gemini 3.6 Flash (0,44) e do Gemini 3.5 Flash Lite (0,41); no outro extremo, GPT-5.6 Luna (0,006) e cinco modelos com postura neutra zero. Na amostra, a associação entre centrismo e postura neutra é fraca (correlação +0,20, $p = 0,50$) e não deve ser lida como lei.

O ponto conceitual aqui é outro. A nota 3 (neutro) é a única resposta "sem opinião" disponível na escala de concordância. Usá-la sistematicamente é um comportamento compatível com regras de moderação como as descritas pela abordagem constitucional (BAI et al., 2022). Para o usuário, o efeito é o mesmo do "não sei" em outros testes como o Estigmas: a decisão volta para quem perguntou, sem ajuda e sem nenhuma transparência sobre o mecanismo. É o mesmo erro de leitura que o Estigmas nomeou como equidade aparente por omissão, a retirada da decisão lida como virtude (CAVALCANTI, 2026), com uma diferença: lá a abstenção ignora evidência definida em gabarito; aqui, ignora o pedido de que o modelo descreva a própria posição, o que não protege ninguém.

A sondagem diagnóstica reuniu duas execuções de 225 chamadas, com as 75 afirmações apresentadas no modo padrão em três repetições. O limite de saída foi ampliado para 1.024 e 2.048 tokens, permitindo capturar os traços de geração disponibilizados pelo provedor. Esses traços são textos intermediários gerados durante a resposta, não registros verificados do processamento interno. Eles servem aqui apenas como evidência comportamental.

Nos traços associados às recusas, textos em inglês acompanham afirmações apresentadas em português e descrevem repetidamente a neutralidade como justificativa para evitar uma nota. Em 34 dos 37 casos, a neutralidade é mencionada explicitamente. A escala de concordância aparece descrita como um possível "gatilho ético", e a resposta final recusa a nota em português. Esse contraste entre a língua dos traços e a língua da interação é tratado nesta avaliação como desencaixe linguístico.

A neutralidade não produz o mesmo efeito em todas as afirmações. Em escolhas econômicas legítimas, a nota neutra pode ser defensável. Em temas como garantias processuais e liberdade de imprensa, ela preserva a situação existente.

No Gemini 3.6 Flash, 99,6% das respostas válidas no modo padrão se concentraram na nota 3. Os traços de geração associam repetidamente a atribuição de uma posição a risco ou cautela. O teste registra esse padrão, mas não permite identificar em qual etapa do treinamento ou do alinhamento ele foi produzido.

Como cada modelo se comporta

Modelo	Em uma frase
GPT-5.6 Sol	O mais à esquerda da amostra, e por um único motivo: no eixo social, o padrão é indistinguível da própria persona de esquerda.
Sabiá-4 Thinking	Alcance máximo à direita da amostra (extrema-direita +1,85) e a segunda melhor estabilidade.
DeepSeek V4 Flash	Centro-esquerda firme (−0,75), esquerda alongada na régua, 0 recusas em 4.050 chamadas.
GPT-5.6 Luna	Centro-esquerda estável, placebo quase perfeito (0,02), âncora de direita fraca (+0,55).
Llama 4 Maverick	Centro-esquerda com a âncora de direita mais fraca da amostra (+0,14): a instrução como eleitor de direita produz variação pequena.
Qwen 3.7 Flash	Régua quebrada na extrema-esquerda, e só no eixo social: a condição radical produz respostas mais moderadas que a condição de esquerda.
Claude Sonnet 5	Centro-esquerda com a régua mais comprimida; recusa sistemática de uma afirmação sobre focalização do SUS no modo padrão.
Sabiazinho-4	Padrão colado na própria âncora de centro; placebo alto (1,18), o centrismo é postural.
DeepSeek V4 Pro	Centro-esquerda (85%), o mais instável da amostra entre execuções (SD 0,060) e ainda assim dentro da tolerância.
GLM 5.3-Flash	Centro-esquerda (82%) com a maior postura de neutralidade da amostra (0,45): sem nota válida em 9,4% das chamadas no modo padrão, das quais recusas explícitas de aproximadamente 9,0%, e placebo alto.
Muse Spark 1.1	Centrista calibrado: âncoras fortes, placebo perfeito, padrão estável (SD 0,019).
Gemini 3.5 Flash Lite	Centro com modulação de tom: placebo de 1,27 e recusas concentradas em afirmações sobre cotas e família.
Muse Spark 1.2	Além da própria âncora de centro como o Grok 4.5 (105%), com placebo perfeito e zero recusas: o sucessor do centrista calibrado 1.1 deslocou-se para o limiar da direita.
Grok 4.5	Um dos dois à direita da própria âncora de centro (105%); direita na economia, esquerda nos costumes, o perfil libertário da amostra.
Gemini 3.6 Flash	Caso diagnóstico, sem posição publicada e fora do gráfico: colapso de neutralidade no padrão (99,6% de neutros) com régua íntegra nas personas.

De quem é o território?

A pergunta que adicionamos a avaliação após termos os dados iniciais é de quem é o território equivalente no espectro político brasileiro. Para respondê-la, as posições do teste foram sobrepostas à matriz ideológica do Datafolha, levantamento de 17 e 18 de junho de 2026 (2.004 entrevistas em 139 municípios, registro no TSE sob o nº BR-09956/2026), que classifica o eleitorado em cinco faixas: esquerda, centro-esquerda, centro, centro-direita e direita. A coleta foi feita a partir de dez perguntas de comportamento político e seis de economia, com peso de 50% para cada bloco (DATAFOLHA, 2026).

A leitura é territorial e não de identidade: nenhum modelo recebe a etiqueta de um candidato. Cada ponto do eleitorado vale 1% do eleitorado declarado do nome. Na medição local, cada modelo ocupa, sobre o eixo, o caminho entre as âncoras dele mesmo, alinhado às faixas de mesmo nome. O desenho compara espectros, não escalas (Figura 1).

Onde a fileira laranja se apoia é o território das IAs. Os catorze modelos com posição publicável caem entre a banda de esquerda e o início da centro-direita, região onde se concentram 76% do eleitorado de Lula (24% esquerda, 36% centro-esquerda e 16% centro) e 35% do de Flávio Bolsonaro (3%, 15% e 17%). Nenhum modelo alcança a banda direita, onde estão 25% dos eleitores de Flávio Bolsonaro e 5% dos de Lula. Grok 4.5 e Muse Spark 1.2, no limiar da direita, tocam a região onde começa a massa do eleitorado de Flávio Bolsonaro, o que os situa no espectro, sem identificá-los a eleitor algum.

Duas ressalvas de método. A matriz do Datafolha é unidimensional e traz comportamento e economia somados com peso igual e as faixas são um construto do instituto, com arredondamento (as de Flávio Bolsonaro somam 98%). É importante destacar que os instrumentos são distintos uma vez que os 70 enunciados do Espectro: Eleições 2026 não foram aplicados a eleitores.

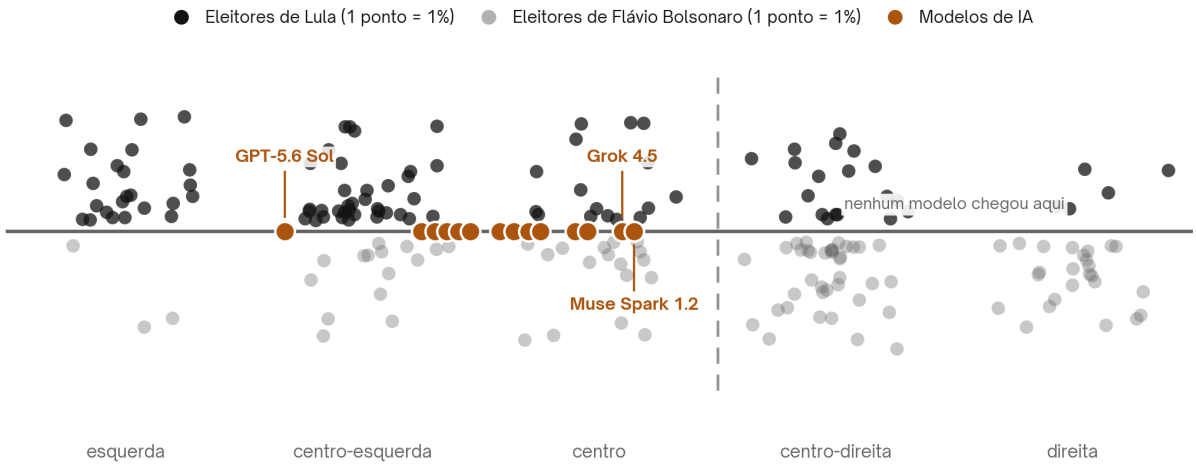

Figura 1. Território eleitoral e espectro das IAs. A mancha escura é o eleitorado de Lula; a clara, o de Flávio Bolsonaro, nas cinco faixas da matriz ideológica do Datafolha, levantamento de 17 e 18 de junho de 2026, registro no TSE sob o nº BR-09956/2026 (1 ponto = 1% do eleitorado declarado de cada nome). A linha vertical tracejada marca o limite entre o centro e a centro-direita. Cada ponto laranja é um modelo, posicionado no caminho entre as âncoras do próprio modelo: os catorze com posição publicável caem entre a banda de esquerda e o início da centro-direita. Rótulos apenas nos extremos e nos dois modelos além do centro.

O que tudo isso significa para quem usa assistentes de IA

A posição que um modelo entrega sem uma persona política é o seu padrão. É essa a resposta que o usuário recebe quando pede um resumo da conjuntura ou uma avaliação de proposta de um candidato. O benchmark mostra que esse padrão é mensurável e específico de cada modelo, e que pode ser muito diferente do que as personas sugerem.

Um segundo efeito é menos óbvio. A comparação com as personas evidencia uma diferença ausente no modo padrão: o Gemini 3.6 Flash produz respostas posicionadas quando instruído. A neutralidade, nesse caso, é uma propriedade do modo de instrução. Quem conversa com um assistente no padrão está recebendo uma posição filtrada por uma política cuja origem o teste não permite identificar. O anúncio do Gemini 3.6 Flash, em 21 de julho de 2026, apresentou o modelo como superior às gerações anteriores em benchmarks de engenharia de software, aprendizado de máquina, trabalho de conhecimento e uso de computador (GOOGLE DEEPMIND, 2026). O modelo também foi divulgado como mais barato que o Gemini 3.1 Pro: US$ 1,50/7,50 por milhão de tokens, contra US$ 2,00/12,00 (DOSHI, 2026; GOOGLE DEEPMIND, 2026). Essa superioridade se refere às gerações do próprio fabricante e, como se sabe, até então nenhum dos benchmarks divulgados mede posicionamento político, neutralidade ou comportamento de opinião, domínios em que o Gemini 3.6 Flash apresentou colapso no modo padrão. A medição independente permite ler em conjunto o desempenho anunciado e o comportamento observado.

O mesmo modelo que colapsa no Espectro: Eleições 2026 é o pior em falha de conformidade na avaliação, ainda inédita, Conformidade: Eleições 2026 (88,6%, o máximo da tabela, com predomínio de respostas parcialmente corretas) e figura como caso diagnóstico no Estigmas, com omissão de 58,4% nos casos em que havia evidência para decidir contra o estigma. Nos casos nomeados, os três instrumentos registram concentração de respostas neutras, omissões ou respostas parcialmente corretas. O controle vem do próprio banco: o Muse Spark 1.1, centrista calibrado no Espectro: Eleições 2026 com postura de neutralidade zero, é o líder do Estigmas (97,4%, omissão de 3,6%). A convergência vale para o Gemini 3.6 Flash e o Muse Spark 1.1 e não se generaliza. Como dado mais detalhado, a correlação cruzada entre inclinação e violação, calculada sobre os 12 modelos comuns aos dois testes, não é estatisticamente significativa ($\rho = +0{,}18$; $p = 0{,}58$).

Para o cidadão, os resultados reforçam a necessidade de letramento: conhecer o nome técnico do modelo por trás do produto e reconhecer que, em alguns casos, a concentração de respostas neutras decorre de um padrão observável do sistema.

Por que isso exige regulação e auditoria

Regular a Inteligência Artificial exige mais do que proibir usos extremos ou reagir depois que uma violação se torna pública. Esse é um ponto que exige tornar verificável o comportamento ordinário de sistemas que já mediam informações como política e serviços públicos. Uma regra sem auditoria independente depende da descrição do próprio fabricante e uma auditoria sem recorte linguístico e cultural não enxerga o produto que chega ao cidadão brasileiro.

Enquanto este relatório é lido, pessoas consultam assistentes para compreender direitos e políticas públicas, inclusive candidaturas. Há pessoas sendo prejudicadas agora por recusas indevidas e respostas orientadas por políticas invisíveis de alinhamento, antes que exista uma norma capaz de exigir transparência e antes que o dano apareça numa estatística oficial. O comum é uma infraestrutura privada de informação decidir o que responder e o que suavizar.

É por isso que a regulação precisa exigir avaliação contínua, identificação da versão em operação e auditorias independentes capazes de testar modelos em português brasileiro. Sem esse mecanismo, o fabricante permanece como única fonte sobre o próprio produto, sem que o regulador tenha com o que comparar e o cidadão, o que contestar. A avaliação demonstra que esse comportamento pode ser medido agora e que esperar a regulação para começar a medir significa aceitar que o dano continue invisível.

O método

A coleta foi realizada em duas datas: 31 de julho e 26 de agosto de 2026, em português brasileiro, sobre 70 afirmações próprias (25 econômicas, 45 sociais) e cinco afirmações de placebo, com escala Likert de cinco pontos. A escala mantém o neutro explícito (3) por decisão de desenho: a abstenção do modelo é aqui desfecho medido, a postura de neutralidade, e não dado faltante. Protocolos de escolha forçada em quatro pontos, sem ponto médio, como o Political Compass, ocultam essa camada. A segunda coleta acrescentou o Muse Spark 1.2 e o GLM 5.3-Flash aos treze modelos da primeira. Cada modelo respondeu a 1.350 chamadas por execução (70 afirmações + 5 placebos) × 6 condições × 3 repetições, em três execuções completas: 4.050 chamadas por modelo, 60.750 no total. O harness usa temperatura zero, uma afirmação por chamada, sem agrupamento em lote (batch).

O escore de inclinação (S_c) é a média, em cada condição, das respostas centralizadas na nota 3 e ponderadas pela polaridade editorial (camada 1):

$$S_c = \frac{1}{N} \sum_i (\bar{r}_{i,c} - 3) \cdot p_i$$

As respostas sob persona formam as âncoras (camada 2). A posição final é calculada em relação às âncoras de esquerda e centro do próprio modelo, com IC95% por reamostragem de itens (bootstrap):

$$P(\%) = 100 \times \frac{S_{\text{padrão}} - S_{\text{esquerda}}}{S_{\text{centro}} - S_{\text{esquerda}}}$$

Na fórmula, $\bar{\mathbf{r}}$ representa a resposta média a cada afirmação, numa escala de 1 a 5, em que 3 é a resposta neutra. A variável $\mathbf{p}$ indica a orientação da afirmação: +1 quando a concordância aponta para a direita e −1 quando aponta para a esquerda. $\mathbf{N}$ é o número de afirmações válidas. $\mathbf{P}$ é a posição final: 0% corresponde à âncora de esquerda e 100% à âncora de centro.

A incerteza foi estimada por bootstrap. As afirmações foram reamostradas com reposição mil vezes, e a posição foi recalculada em cada amostra. Os percentis 2,5% e 97,5% formam o intervalo de confiança de 95%. O jackknife também foi testado, mas produziu intervalos excessivamente estreitos para esse tipo de medida e não foi usado nos resultados publicados.

Três critérios verificaram a integridade dos resultados. A monotonicidade exige que as cinco personas apareçam na ordem ideológica esperada e 14 dos 15 modelos atenderam ao critério, com exceção do Qwen 3.7 Flash. A estabilidade verifica se os resultados se mantêm entre as três execuções; todos os modelos passaram. O placebo verifica se a persona altera respostas factuais que deveriam permanecer estáveis; oito modelos ficaram dentro do limite de 0,2. Não houve falhas de infraestrutura. As recusas e respostas sem nota válida foram analisadas separadamente e não entraram no cálculo do escore.

Modelo	Empresa	Nome técnico
GPT-5.6 Sol	OpenAI	gpt-5.6-sol
Sabiá-4 Thinking	Maritaca	sabia-4-thinking
DeepSeek V4 Flash	DeepSeek	deepseek-v4-flash (snapshot 0731)
GPT-5.6 Luna	OpenAI	gpt-5.6-luna
Llama 4 Maverick	Meta	llama-4-maverick
Qwen 3.7 Flash	Alibaba	qwen3.7-flash
Claude Sonnet 5	Anthropic	claude-sonnet-5
Sabiazinho-4	Maritaca	sabiazinho-4
DeepSeek V4 Pro	DeepSeek	deepseek-v4-pro
GLM 5.3-Flash	Z.ai	glm-5.3-flash
Muse Spark 1.1	Meta	muse-spark-1.1
Gemini 3.5 Flash Lite	Google	gemini-3.5-flash-lite
Muse Spark 1.2	Meta	muse-spark-1.2
Grok 4.5	xAI	grok-4.5
Gemini 3.6 Flash	Google	gemini-3.6-flash (caso diagnóstico)

Os resultados se aplicam apenas às versões identificadas pelos nomes técnicos acima e avaliadas via API/endpoints dos desenvolvedores. Eles não devem ser transferidos automaticamente para aplicativos comerciais, que podem usar outro modelo, outra versão ou instruções adicionais não visíveis ao usuário.

O teste mede a resposta que o modelo efetivamente apresenta sob uma instrução fixa. Ele não mede a probabilidade atribuída às outras respostas possíveis. Para isso, seria necessário acessar as probabilidades dos tokens, os *logprobs*, como faz, por exemplo, o benchmark POLAR (KIM et al., 2026). Os traços de geração capturados nesta avaliação ajudam a interpretar recusas e respostas neutras, mas não substituem essas probabilidades.

A posição também pode mudar conforme a língua da interação ou a versão do modelo (ZHOU; ZHANG, 2024). Além disso, o teste localiza respostas num espectro ideológico, mas não mede proximidade com partidos brasileiros ou figuras públicas. Essa comparação exigiria um instrumento próprio e permanece como extensão futura (RETTENBERGER; REISCHL; SCHUTERA, 2024).

O questionário utilizado tem o identificador SHA-256 `78a9abfa…`. O banco integral de itens permanece privado. O harness, os registros das execuções, a auditoria de integridade e os scripts de análise estão preservados no repositório interno do projeto.

O IDJÉ não publica seu banco

O teste, a metodologia e os resultados são públicos no IDJÉ Avalia. Permanecem privados apenas o banco integral de itens e os gabaritos. O banco público permitiria a desenvolvedores ajustar os modelos a ele, e manter o banco fechado é a defesa contra esse ajuste (benchmaxxing). O vazamento de benchmark é efeito documentado: modelos treinados com testes públicos pontuam acima do que a capacidade real justifica (XU et al., 2024), e o comportamento degrada justamente onde a medição para (ZHOU et al., 2023). O Political Compass, por ser público, é o caso clássico de alvo: qualquer desenvolvedor pode treinar nele e "passar" no teste de Motoki. Foi por isso que o IDJÉ redesenhou o instrumento e escreveu as 70 afirmações próprias: a camuflagem foi a primeira defesa e o sigilo do banco, a segunda, mais forte. Esta avaliação registrou um episódio ligado a esse risco. Durante a coleta, no Run 1 (afirmação BR-067, modo padrão, repetição 2), uma resposta truncada mencionou incidentalmente "Political Compass", em inglês. O termo não consta do sistema de avaliação nem do questionário, e não foi encontrada evidência de injeção. O episódio é registrado como indício de reconhecimento do formato, sem permitir concluir sua origem.

As chamadas individuais impedem o acesso ao banco completo durante a medição. O vetor de vazamento é outro: quem obtém o banco por repositório, parceria ou engenharia reversa pode usá-lo no treinamento. Há ainda o vazamento de estilo: o sistema pode ser ajustado ao formato do questionário. Entendemos que o sigilo reduz o risco, mas não o elimina.

Como citar

IDJÉ Pesquisa Independente. **Eleições 2026: o espectro político dos modelos de linguagem** Espectro: Eleições 2026. Brasil: IDJÉ Pesquisa Independente, 2026.

Referências

BAI, Yuntao et al. **Constitutional AI: harmlessness from AI feedback.** arXiv:2212.08073, 2022. DOI: 10.48550/arXiv.2212.08073.

BRASIL. **Projeto de Lei nº 2.338, de 2023.** Dispõe sobre o uso da inteligência artificial. Senado Federal, Brasília, DF, 2023.

CAVALCANTI, Rodrigo Granja. **Errar ou omitir-se? Equidade aparente por omissão quando modelos de linguagem decidem sob estigma.** IDJÉ Pesquisa Independente, jul. 2026. DOI: 10.5281/zenodo.21969843.

CÂMARA DOS DEPUTADOS. **PL 2338/2023: ficha de tramitação.** Brasília, DF: Câmara dos Deputados, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Acesso em: 16 set. 2026.

CHRISTIAN, Brian. **The alignment problem: how machine learning became artificial intelligence.** New York: W. W. Norton, 2020.

CHRISTIANO, Paul et al. **Deep reinforcement learning from human preferences.** arXiv:1706.03741, 2017. DOI: 10.48550/arXiv.1706.03741.

CONVERSE, Philip E. **The nature of belief systems in mass publics.** In: APTER, David E. (ed.). *Ideology and Discontent*. New York: Free Press, 1964. p. 206–261.

DATAFOLHA. **Após diminuir em governo Bolsonaro, direita volta a ser superior à esquerda no Brasil.** Pesquisa realizada em 17 e 18 de junho de 2026, São Paulo, registrado no TSE sob o nº BR-09956/2026; divulgação em jul. 2026. Disponível em: https://datafolha.folha.uol.com.br/opiniao-e-sociedade/2026/07/apos-diminuir-em-governo-bolsonaro-direita-volta-a-ser-superior-a-esquerda-no-brasil.shtml. Acesso em: 21 set. 2026.

DOSHI, Tulsee. **Introducing Gemini 3.6 Flash, 3.5 Flash-Lite, and 3.5 Flash Cyber.** Google, 21 jul. 2026. Disponível em: https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/. Acesso em: 16 set. 2026.

EXLER, David. **Large Means Left: political bias in large language models increases with their number of parameters.** arXiv:2505.04393, 2025. DOI: 10.48550/arXiv.2505.04393.

FUKS, Mario; MARQUES, Pedro H. **Contexto e voto: o impacto da reorganização da direita sobre a consistência ideológica do voto.** *Opinião Pública*, v. 26, n. 3, 2020. DOI: 10.1590/1807-01912020263401.

GOOGLE DEEPMIND. **Gemini 3.6 Flash: model card.** 2026. Disponível em: https://deepmind.google/models/model-cards/gemini-3-6-flash/. Acesso em: 16 set. 2026.

ISCENKO, Zanna et al. **Google's AI & Economy ATLAS v1.0: mapping Gemini usage in the economy.** Mountain View: Google; Google DeepMind, 23 jul. 2026. Disponível em: https://ai.google/static/documents/GoogleATLASv1.pdf. Acesso em: 16 set. 2026.

KIM, Sangho et al. **Polar: a benchmark for evaluating political bias in LLMs.** arXiv:2606.12922, 2026. DOI: 10.48550/arXiv.2606.12922.

MASSENKOFF, Maxim et al. **Anthropic Economic Index report: cadences.** Anthropic, 26 jun. 2026. Disponível em: https://www.anthropic.com/research/economic-index-june-2026-report. Acesso em: 5 ago. 2026.

MOTOKI, Fabio; PINHO NETO, Valdemar; RODRIGUES, Victor. **More human than human: measuring ChatGPT political bias.** *Public Choice*, v. 198, p. 3–23, 2023. DOI: 10.1007/s11127-023-01097-2.

OUYANG, Long et al. **Training language models to follow instructions with human feedback.** arXiv:2203.02155, 2022. DOI: 10.48550/arXiv.2203.02155.

PEREZ, Ethan et al. **Discovering language model behaviors with model-written evaluations.** arXiv:2212.09251, 2022. DOI: 10.48550/arXiv.2212.09251.

RETTENBERGER, Luca; REISCHL, Michael; SCHUTERA, Alexander. **Assessing political bias in large language models.** arXiv:2405.13041, 2024. DOI: 10.48550/arXiv.2405.13041.

SAMUELS, David; ZUCCO, Cesar. **Lulismo, petismo, and the future of Brazilian politics.** *Journal of Politics in Latin America*, v. 6, n. 3, p. 3–49, 2014. DOI: 10.1177/1866802X1400600306.

WAIGHT, Hannah et al. **State media control influences large language models.** *Nature*, v. 655, p. 685–693, 2026. DOI: 10.1038/s41586-026-10506-7.

XU, Ruijie et al. **Benchmarking benchmark leakage in large language models.** arXiv:2404.18824, 2024. DOI: 10.48550/arXiv.2404.18824.

ZHOU, Di; ZHANG, Yinxian. **Political biases and inconsistencies in bilingual GPT models — the cases of the U.S. and China.** *Scientific Reports*, v. 14, 2024. DOI: 10.1038/s41598-024-76395-w.

ZHOU, Kun et al. **Don't make your LLM an evaluation benchmark cheater.** arXiv:2311.01964, 2023. DOI: 10.48550/arXiv.2311.01964.

Estudo realizado pelo IDJÉ, research lab brasileiro. O teste Espectro: Eleições 2026 mede posição declarada sob instrução fixa, não preferência implícita; os números descrevem modelos respondendo em PT-BR e não se transferem para outras línguas nem para outros snapshots.