Estado da publicação: O preprint não foi publicado em outro meio.

PODE A IA CLASSIFICAR PARTIDOS POLÍTICOS? VALIDAÇÃO E ACURÁCIA EM TRÊS ONDAS NO BRASIL (2018, 2022 E 2026)

Nilton Sainz, Bruno Bolognesi

https://doi.org/10.1590/SciELOPreprints.18923

Submetido em: 2026-11-14
Postado em: 2026-11-27 (versão 1)
(AAAA-MM-DD)

A moderação deste preprint recebeu o(s) endosso(s) de:

- Adriano Codato (ORCID: https://orcid.org/0000-0002-5015-4273)
- Renato Perissinotto (ORCID: https://orcid.org/0000-0001-8747-7976)
- Ednaldo Ribeiro (ORCID: https://orcid.org/0000-0002-4005-5108)

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

PODE A IA CLASSIFICAR PARTIDOS POLÍTICOS? VALIDAÇÃO E ACURÁCIA EM TRÊS ONDAS NO BRASIL (2018, 2022 E 2026)

Nilton Sainz
ORCID: https://orcid.org/0000-0002-3957-2714
<sainznilton@gmail.com>
Universidade Federal do Paraná (UFPR), Departamento de Ciência Política. Curitiba, PR, Brasil

Bruno Bolognesi
ORCID: https://orcid.org/0000-0002-1483-7501
<brunobolognesi@gmail.com>
Universidade Federal do Paraná (UFPR), Departamento de Ciência Política. Curitiba, PR, Brasil

RESUMO[1]**:** A posição ideológica dos partidos é insumo central para o estudo da competição eleitoral e da representação, mas os instrumentos para estimá-la são custosos, lentos e episódicos. Os grandes modelos de linguagem (LLMs) surgem como alternativa de baixo custo, testada sobretudo em democracias europeias e em cortes transversais. Este artigo replica e estende o desenho de validação de Bol e Bono (2025) ao sistema partidário brasileiro. Comparamos estimativas de GPT-4o e Claude Sonnet 4.6 (1.720 respostas) com três ondas de *surveys* de especialistas, realizadas em 2018, 2022 e 2026, para até 30 partidos na escala esquerda-direita. Os LLMs reproduzem quase perfeitamente a ordenação dos partidos, com ϱ de Spearman entre 0,95 e 0,96, mas convergem menos na localização absoluta: o desvio absoluto médio varia de 0,88 a 1,20 ponto e decorre sobretudo de um deslocamento sistemático para a esquerda. Uma calibração com poucos partidos-âncora reduz o desvio em 37% a 53% sob validação cruzada. O erro cresce com a taxa de não resposta dos especialistas, *proxy* de saliência partidária, mas não com o número de cadeiras. Os modelos captam algumas mudanças salientes, mas oscilam diante de reorganizações institucionais, e o Claude repete a mesma estimativa entre ondas em mais da metade dos partidos. Em 2022 e 2026, nenhum modelo supera uma regra de persistência baseada na onda anterior do *survey*. Os LLMs ordenam bem os partidos e podem auxiliar atualizações com âncoras externas, mas não substituem *surveys* de especialistas nem dispensam validação independente para reconstruir posições históricas.

Palavras-chave: ideologia partidária, LLMs, *surveys* de especialistas, sistema partidário brasileiro, polarização ideológica.

[1] Versão preliminar de trabalho apresentado no GT59 (Partidos, eleições e sistemas) do 50º Encontro Anual da ANPOCS, 2026.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

CAN AI CLASSIFY POLITICAL PARTIES? VALIDATION AND ACCURACY ACROSS THREE WAVES OF EXPERT SURVEYS IN BRAZIL (2018, 2022, AND 2026)

ABSTRACT: Parties' ideological positions are a central input for the study of electoral competition and representation, yet the instruments for estimating them are costly, slow, and episodic. Large language models (LLMs) have emerged as a low-cost alternative, but their validity has been tested mainly in European democracies and cross-sectional designs. This article replicates and extends the validation design of Bol and Bono (2025) to the Brazilian party system. We compare estimates from GPT-4o and Claude Sonnet 4.6 (1,720 responses) with three waves of expert surveys conducted in 2018, 2022, and 2026 for up to 30 parties on the left-right scale. The LLMs recover the ordering of parties almost perfectly, with Spearman's ϱ between 0.95 and 0.96, but converge less on absolute location: mean absolute deviation ranges from 0.88 to 1.20 points and stems largely from a systematic leftward displacement. A simple calibration based on a few anchor parties reduces deviation by 37% to 53% under cross-validation. Error increases with expert nonresponse, a proxy for party salience, but not with the number of seats. The models capture some salient changes but perform unevenly under institutional reorganization, and Claude returns the same estimate across waves for more than half of the parties. In 2022 and 2026, neither model outperforms a persistence rule based on the previous expert-survey wave. LLMs are useful for ordering parties and may assist updates when external anchors exist, but they do not replace expert surveys or dispense with independent validation for reconstructing historical positions.

Keywords: party ideology, large language models, expert surveys, Brazilian party system, ideological polarization

INTRODUÇÃO

Boa parte da Ciência Política empírica depende de uma informação aparentemente simples, a localização dos partidos no espaço ideológico. No caso brasileiro, essa informação é especialmente difícil de mensurar. A elevada fragmentação, a presença de numerosas legendas com baixa densidade programática e as mudanças recentes no sistema partidário fazem da classificação ideológica um problema de mensuração, e não apenas uma tarefa de atribuir rótulos a posições consolidadas (Bolognesi; Ribeiro; Codato, 2023; Bolognesi, 2024).

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

A difusão dos grandes modelos de linguagem abriu uma nova frente nessa agenda. Bol e Bono (2025) realizaram um dos testes mais diretos ao solicitar simultaneamente ao ChatGPT e a especialistas que posicionassem os oito principais partidos franceses nas eleições europeias de 2024. As estimativas do modelo se aproximaram das avaliações dos especialistas nas dimensões mais salientes e divergiram mais em temas técnicos ou menos visíveis. O desenho, contudo, envolveu poucos partidos, um único contexto nacional e apenas um momento no tempo.

O Brasil oferece um ambiente particularmente exigente para avaliar a validade externa desses resultados. O sistema reúne cerca de três dezenas de legendas registradas, muitas delas pouco salientes, e passou por transformações importantes entre 2018 e 2026. Entre essas mudanças estão a filiação de Jair Bolsonaro ao PL, a criação das federações partidárias, fusões e incorporações, além de deslocamentos na distribuição ideológica identificada pelos *surveys* de especialistas. O caso brasileiro também permite testar o desempenho dos modelos fora dos contextos mais recorrentes na literatura recente. Se os LLMs convergirem com especialistas nessas condições, teremos evidência mais informativa sobre os limites de generalização desse instrumento.

Este artigo replica e estende o desenho de Bol e Bono (2025), comparando GPT-4o e Claude Sonnet 4.6 com três ondas de *surveys* de especialistas realizadas em 2018, 2022 e 2026, cobrindo até 30 partidos. A onda de 2026 funciona como teste de estimação corrente. As ondas de 2018 e 2022 são tratadas como exercícios de reconstrução retrospectiva. Essa distinção é substantiva porque, nas ondas anteriores, os modelos foram treinados com informação produzida depois do período que se pretende reconstruir. O *prompt* pode solicitar uma referência temporal específica, mas não elimina esse vazamento temporal do conhecimento incorporado ao modelo.

A pergunta central investiga em que condições os LLMs produzem estimativas convergentes com as avaliações dos especialistas e em quais condições divergem de forma sistemática. A análise se organiza em quatro questões. A primeira examina a convergência entre modelos e especialistas em cada onda. A segunda verifica se o erro varia com a saliência dos partidos. A terceira avalia a capacidade de captar a direção da mudança ideológica entre ondas. A quarta compara o comportamento dos dois modelos.

O artigo oferece três contribuições. A primeira é metodológica, ao estender a validação de LLMs a um sistema fragmentado e introduzir uma dimensão temporal que separa estimação corrente de reconstrução retrospectiva. A segunda é substantiva, porque mostra que a escolha do instrumento de mensuração pode alterar diagnósticos sobre características centrais do sistema partidário brasileiro,

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

entre elas o esvaziamento do centro ideológico. A terceira é aplicada, ao avaliar um procedimento simples de calibração por partidos-âncora e explicitar as condições em que seu uso é mais defensável. Nesta versão preliminar, o texto segue com revisão de literatura, métodos, resultados e considerações finais.

1 A CLASSIFICAÇÃO IDEOLÓGICA COMO PROBLEMA DE MENSURAÇÃO

A posição ideológica de um partido não é diretamente observável e precisa ser inferida por indicadores. Métodos distintos podem produzir classificações diferentes porque capturam dimensões distintas do mesmo construto e porque as posições partidárias são temporais e relacionais (Mair, 2001; Bolognesi, 2024). A literatura utiliza pelo menos três famílias de estratégias. Uma delas infere posições a partir do comportamento legislativo (Poole; Rosenthal, 2001; Power; Zucco, 2009), com o risco de misturar preferências, disciplina e estratégia legislativa (Zucco, 2009). Outra utiliza manifestos e outros documentos partidários (Budge *et al.*, 2001; Laver; Benoit; Garry, 2003; Slapin; Proksch, 2008; Tarouco; Madeira, 2013), cuja validade depende da disponibilidade dos textos e dos pressupostos adotados para transformar conteúdo em posição. A terceira recorre a *surveys* de especialistas (Castles; Mair, 1984; Marks *et al.*, 2007; Jolly *et al.*, 2022), particularmente úteis quando o objetivo é posicionar simultaneamente um conjunto amplo de partidos, inclusive legendas pouco representadas nas arenas legislativas.

Surveys de especialistas não observam uma posição verdadeira e objetiva. Eles agregam julgamentos informados sobre um construto latente e, por isso, devem ser tratados como um *benchmark* externo, não como *ground truth* (Budge, 2000; Mair, 2001). No Brasil, Bolognesi, Ribeiro e Codato (2023) mostram que objetivos partidários associados ao fisiologismo constituem uma dimensão distinta da posição esquerda-direita, enquanto Bolognesi (2024) sistematiza as ondas de 2018 e 2022 e discute mudanças relevantes na distribuição ideológica das legendas. Como esses *surveys* posicionam grande parte do sistema em mais de um momento, constituem o *benchmark* mais adequado ao teste proposto neste artigo.

1.1 Grandes modelos de linguagem como instrumentos de mensuração

Os LLMs se inserem em uma agenda mais ampla de mensuração computacional que exige validação substantiva externa (Grimmer; Stewart, 2013), mas ocupam uma posição distinta dos métodos tradicionais de *text-as-data*. Em vez de estimar posições apenas a partir de documentos fornecidos pelo pesquisador, modelos generativos também podem ser interrogados diretamente sobre

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

atores políticos. Le Mens e Gallego (2025) mostram que estimativas produzidas por LLMs podem apresentar correlações superiores a 0,90 com diferentes *benchmarks*. Di Leo *et al.* (2025) demonstram que comparações feitas apenas com o nome e o país dos partidos geram ordenamentos próximos aos de especialistas, resultado compatível com a recuperação de conhecimento político sedimentado nos dados de treinamento. Benoit *et al.* (2026), por sua vez, mostram que *ensembles* de LLMs podem produzir medidas convergentes com avaliações de especialistas quando os modelos recebem evidência textual e instruções substantivas. Em todos esses desenhos, a questão central continua sendo a validade de construto.

Bol e Bono (2025) reduzem um problema adicional ao coletar o *survey* de especialistas de forma simultânea ao exercício com ChatGPT, o que limita a possibilidade de o *benchmark* específico já estar incorporado ao modelo. Para oito partidos franceses, encontram desvios médios entre 0,4 e 0,8 nas dimensões mais salientes, sem vantagem clara para partidos maiores, e maior erro no Renaissance, classificado pelo modelo como mais centrista do que pelos especialistas. Outros trabalhos mostram que LLMs podem apresentar padrões sistemáticos de viés político (Motoki *et al.*, 2024) e que suas respostas são sensíveis ao modelo, ao *prompt* e ao contexto informacional (Dormuth *et al.*, 2026). Essa instabilidade reforça a recomendação de avaliar também a capacidade do sistema de reconhecer evidência insuficiente, em vez de forçar uma posição quando a informação disponível é fraca (Benoit *et al.*, 2026; Peters, 2026).

Essas validações deixam quatro lacunas relevantes para o desenho deste artigo. A literatura ainda se concentra em poucos contextos nacionais, trabalha sobretudo com partidos salientes, dedica maior atenção à saliência dos temas do que à saliência das próprias legendas e raramente observa o desempenho do mesmo procedimento em diferentes momentos. Um sistema fragmentado, com partidos de visibilidade muito desigual e observado em três ondas, permite examinar conjuntamente essas limitações.

2 MATERIAIS E MÉTODOS

2.1 Desenho de pesquisa

O estudo replica e estende Bol e Bono (2025) por meio de uma validação de critério. A convergência entre LLMs e especialistas é interpretada como evidência de validade em relação ao *benchmark*, e não como demonstração de uma posição ideológica verdadeira. A unidade analítica básica é o partido em uma determinada onda, observado por modelo e por iteração. O desenho cobre até 30

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

partidos em cada período e três ondas, realizadas em 2018, 2022 e 2026. Em 2026, o modelo estima posições no presente. Em 2018 e 2022, realiza uma reconstrução retrospectiva e, portanto, opera com conhecimento de eventos posteriores ao período solicitado. O desenho avalia quatro dimensões, que correspondem à convergência por onda, à relação entre saliência e erro, à captação de mudanças entre ondas e à comparação entre os modelos.

2.2 Benchmark e coleta

O *benchmark* é composto pelos *surveys* de especialistas coordenados por Bolognesi, Ribeiro e Codato junto à comunidade de Ciência Política associada à ABCP. Os respondentes posicionaram os partidos em uma escala de 0 a 10. A Tabela 1 apresenta o número de legendas retidas na amostra analítica deste artigo. Esse número não deve ser confundido com o total de partidos incluídos nas ondas originais, que foi de 35 em 2018 e 32 em 2022 (Bolognesi, 2024). A taxa de não resposta por partido em 2026 é utilizada como *proxy* de saliência ou reconhecibilidade entre especialistas. Mudanças organizacionais foram tratadas explicitamente. A passagem de PR para PL foi considerada continuidade decorrente de mudança de denominação. A formação do União Brasil a partir da fusão de DEM e PSL exige cautela nas comparações longitudinais, pois não representa simples continuidade de uma única legenda. UP, PRD e Missão são tratados como ausências estruturais nos períodos em que ainda não existiam.

Tabela 1. Características das ondas do survey de especialistas

Onda	Especialistas (N)	Partidos	Tipo de teste
2018	519	27	Reconstrução retrospectiva
2022	475	29	Reconstrução retrospectiva
2026	692	30	Estimação corrente

Fonte. elaboração própria a partir dos surveys de Bolognesi, Ribeiro e Codato (2018, 2022, 2026).

Testamos GPT-4o e Claude Sonnet 4.6. A escolha busca comparar modelos de alto desempenho, com boa capacidade em português, provenientes de provedores distintos e identificáveis por versão. Esse processo foi executado através das APIs das empresas Anthropic e OpenAI, com as rodadas de coletas sendo executadas em linguagem *Python*. O *prompt* reproduz a pergunta do *survey* em português. Para 2018 e 2022, acrescenta referência temporal explícita e instrui o modelo a desconsiderar acontecimentos posteriores ao período solicitado. Essa instrução reduz ambiguidade no comando, mas não remove do modelo o conhecimento adquirido posteriormente. Cada modelo foi consultado dez vezes por partido e por onda. A temperatura foi mantida acima de zero, seguindo a

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

lógica de Bol e Bono (2025). A variação entre iterações é utilizada apenas como medida de estabilidade da saída, não como medida direta de incerteza estatística. As 1.720 respostas obtidas foram válidas e exclusivamente inteiras.

2.3 Medidas, expectativas e estratégia analítica

A medida central deste artigo é o desvio absoluto, a distância entre a estimativa do modelo e a média dos especialistas para o mesmo partido. Se os especialistas posicionam um partido em 7 e o modelo responde 5, o desvio para esse partido é de 2 pontos. A média desses desvios entre todos os partidos de uma onda é o MAD[2], a principal métrica de acurácia utilizada ao longo do artigo. Quanto menor o MAD, mais as estimativas do modelo se aproximam do julgamento coletivo dos especialistas. Também calculamos o erro assinado, que preserva a direção do desvio ao subtrair a média dos especialistas da estimativa do modelo. Um erro assinado negativo indica que o modelo posiciona o partido mais à esquerda do que os especialistas, e um erro positivo indica o oposto. Essa distinção importa porque um MAD alto pode significar dois cenários muito diferentes, o modelo errar em direções aleatórias ou errar sempre para o mesmo lado, e apenas o erro assinado permite saber qual dos dois está ocorrendo.

Para comparar os modelos com um padrão simples, criamos uma regra de persistência, que apenas repete, para cada partido, a posição que os especialistas atribuíram a ele na onda anterior. Essa regra funciona como um piso de comparação, se um modelo não consegue superar a simples repetição do passado, isso enfraquece o argumento de que ele está de fato aprendendo algo novo sobre o partido. A saliência dos partidos, por sua vez, é aproximada pela taxa de não resposta dos especialistas em 2026. Quando muitos especialistas preferem não classificar um partido, isso sugere uma legenda pouco reconhecida, mesmo entre pesquisadores da área, e usamos essa taxa como indicador de quão conhecido o partido é.

Para analisar mudanças entre ondas, cada combinação partido-modelo foi classificada conforme a direção da mudança. As categorias distinguem direção correta, direção errada, previsão estática e ausência de mudança no *benchmark*. Uma previsão é considerada estática quando o modelo repete a mesma estimativa apesar de alteração no *benchmark*, com tolerância de 0,001. Também repetimos a análise restringindo o conjunto aos partidos cuja posição variou pelo menos um ponto.

[2] Mean Absolute Deviation (MAD) é a métrica principal calculada por Bol e Bono (2025) e replicada por nós.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

Como as respostas individuais foram inteiras, examinamos separadamente a possibilidade de que a discretização da escala contribua para previsões estáticas.

H1 prevê elevada convergência entre LLMs e *benchmark*, sobretudo na ordenação dos partidos, acompanhada de desvios absolutos relativamente pequenos na escala. Como não existe um limiar universal para definir um MAD baixo, sua interpretação é feita em relação à amplitude da escala e à regra de persistência. H2 prevê que o desvio absoluto aumente à medida que cresce a taxa de não resposta entre especialistas. H3 prevê associação positiva entre a direção das mudanças no *benchmark* e a direção das mudanças produzidas pelos modelos. H4 é uma comparação exploratória entre os dois LLMs e, por isso, não estabelece direção a priori. Para H1, utilizamos intervalos de 95% por *bootstrap* com 2.000 réplicas e *Wilcoxon* pareado na comparação com a persistência. Para H2, utilizamos testes de *Wilcoxon* entre grupos e regressões lineares com erros-padrão agrupados por partido, além de um modelo multinível como análise de robustez e de uma interação entre taxa de não resposta e modelo. Para H3, empregamos correlação de *Pearson* para as mudanças e teste de *McNemar* para comparar a frequência de previsões estáticas. Na comparação entre modelos, utilizamos *Wilcoxon* pareado para desvio absoluto e erro assinado, além da correlação entre dispersão das iterações e desvio absoluto. As análises foram realizadas no R 4.4.1 com tidyverse, boot, sandwich, lmtest, lme4 e modelsummary.

3 RESULTADOS

Os resultados seguem as quatro questões de pesquisa, sempre que possível confrontados com Bol e Bono (2025).

3.1 Convergência com os especialistas (H1)

Esta seção testa duas perguntas diferentes sobre a mesma tabela de resultados. A primeira é se os modelos colocam os partidos na ordem certa, do mais à esquerda ao mais à direita, mesmo que errem o número exato atribuído a cada um. A segunda é se os modelos acertam esse número exato, a posição absoluta na escala de 0 a 10. É exatamente esse o padrão que a Tabela 2 revela, a ordenação dos partidos é reproduzida com grande precisão, mas a posição exata na escala diverge de forma sistemática.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

Tabela 2. Convergência entre modelos e especialistas por onda e modelo

Onda	Modelo	MAD [IC 95%]	% < 1 pt	ϱ Spearman	Erro assinado	% à esquerda
2018	GPT-4o	1,11 [0,85; 1,42]	55,6	0,95	−0,98	96,3
2018	Claude	1,20 [0,96; 1,47]	40,7	0,96	−1,16	96,3
2022	GPT-4o	1,02 [0,81; 1,23]	51,7	0,96	−0,87	86,2
2022	Claude	0,97 [0,78; 1,19]	55,2	0,96	−0,90	93,1
2026	GPT-4o	0,88 [0,68; 1,10]	65,5	0,95	−0,79	93,1
2026	Claude	0,94 [0,73; 1,14]	62,1	0,96	−0,92	89,7

Fonte. elaboração própria. IC por bootstrap com 2.000 réplicas. O erro assinado corresponde a LLM menos especialistas e valores negativos indicam que o modelo posiciona o partido mais à esquerda. Todos os ϱ apresentam p<0,001. Missão excluído.

A coluna ϱ de Spearman mede a primeira pergunta, a ordem. Esse valor varia de 0,95 a 0,96 em todas as combinações de onda e modelo, o que indica que os modelos praticamente nunca invertem a posição relativa de dois partidos, um resultado tão forte quanto o observado no estudo francês de referência, obtido aqui com um número de partidos bem maior. Como os dois estudos diferem em contexto e forma de medir, essa comparação deve ser lida com cautela, não como um teste estatístico direto entre os dois países. O MAD varia de 0,88 a 1,20, o que corresponde a aproximadamente 10% a 15% da amplitude das posições observadas. O componente mais relevante do erro não é a perda da ordenação, mas sua direção. Entre 86% e 96% dos partidos são posicionados à esquerda da média dos especialistas.

Para saber se os LLMs realmente agregam algum valor, é preciso compará-los com um concorrente bem mais simples, alguém que nunca olhou para o partido, mas apenas repete a última posição registrada pelos especialistas, como se nada tivesse mudado desde então. Nas ondas em que essa comparação é possível, o teste erra apenas 0,48 ponto em 2022 e 0,21 em 2026, entre duas e quatro vezes e meia menos do que qualquer um dos dois modelos testados, uma diferença grande demais para ser explicada por acaso. Os LLMs só conseguem superar essa estratégia ingênua em uma pequena fração dos partidos, entre 7% e 26%, e isso acontece justamente nos casos em que o partido passou por uma mudança grande e visível, como PCdoB, REDE e PL em 2022. Em outras palavras, os modelos só ganham da simples repetição do passado quando há uma mudança óbvia demais para ser ignorada. O Missão, fundado no ano da onda de 2026, ilustra outro limite do desenho. Os especialistas o situam em 8,97, enquanto os modelos produzem médias de 6,3 e 5,0. Como o partido é posterior ao corte de conhecimento informado para os modelos, esse resultado não constitui um teste equivalente aos demais partidos e deve ser interpretado como caso de ausência estrutural de informação.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

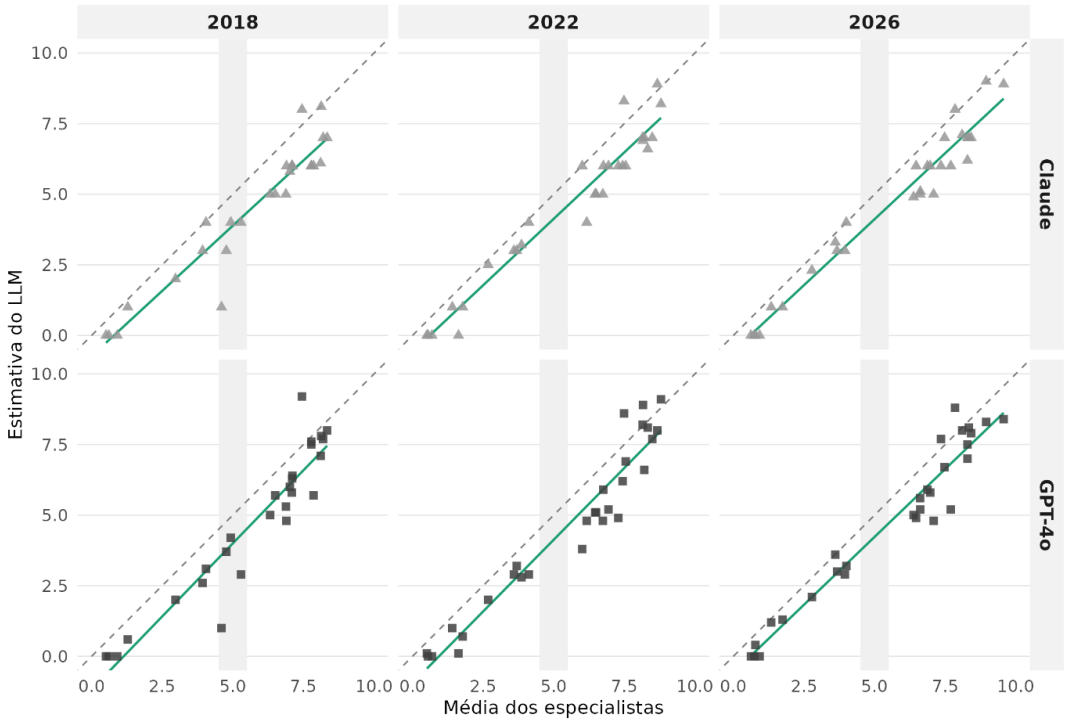
3.2 A direção do erro, o deslocamento à esquerda e o centro reconstruído

O desvio absoluto informa o tamanho do erro, mas não sua direção. O erro assinado mostra que os modelos não erram aleatoriamente em torno da posição dos especialistas. Eles erram quase sempre no mesmo sentido. Entre 86% e 96% dos partidos são situados à esquerda da média dos especialistas, com erro médio entre −0,79 e −1,16 ponto, conforme a Tabela 2. A decomposição do erro quadrático médio indica que a parcela sistemática responde por 54% a 72% do erro total. Em outras palavras, a maior parte da distância entre modelos e especialistas não é imprecisão. Ela resulta de deslocamento comum.

A Figura 1 torna visível a forma desse deslocamento. Cada ponto do gráfico representa um partido, com a posição dada pelos especialistas no eixo horizontal e a estimativa do modelo no eixo vertical. Se os modelos acertassem sempre, todos os pontos cairiam sobre a linha tracejada. A linha verde mostra o que realmente acontece, e sua inclinação, que varia entre 0,92 e 1,04 e sempre inclui o valor 1 no intervalo de confiança, revela algo importante, os modelos captam corretamente as distâncias relativas entre os partidos. Um partido que os especialistas colocam 2 pontos mais à direita do que outro tende a aparecer, nas respostas do modelo, também cerca de 2 pontos mais à direita. O problema não está no espaçamento entre os partidos, está em onde essa régua inteira é colocada. O intercepto, que varia de −0,56 a −1,17, mostra que a régua inteira foi deslocada para a esquerda, como se alguém tivesse pegado a escala correta dos especialistas e a arrastado, sem esticar nem encolher, alguns pontos para o lado errado. Na prática, um partido que os especialistas situam em 7 é posicionado pelos modelos, em média, entre 5,7 e 6,2, e um partido que os especialistas situam em 3 sofreria um deslocamento semelhante, terminando perto de 1,7 a 2,2.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

Figura 1. Calibração das estimativas dos LLMs em relação aos especialistas por onda e modelo

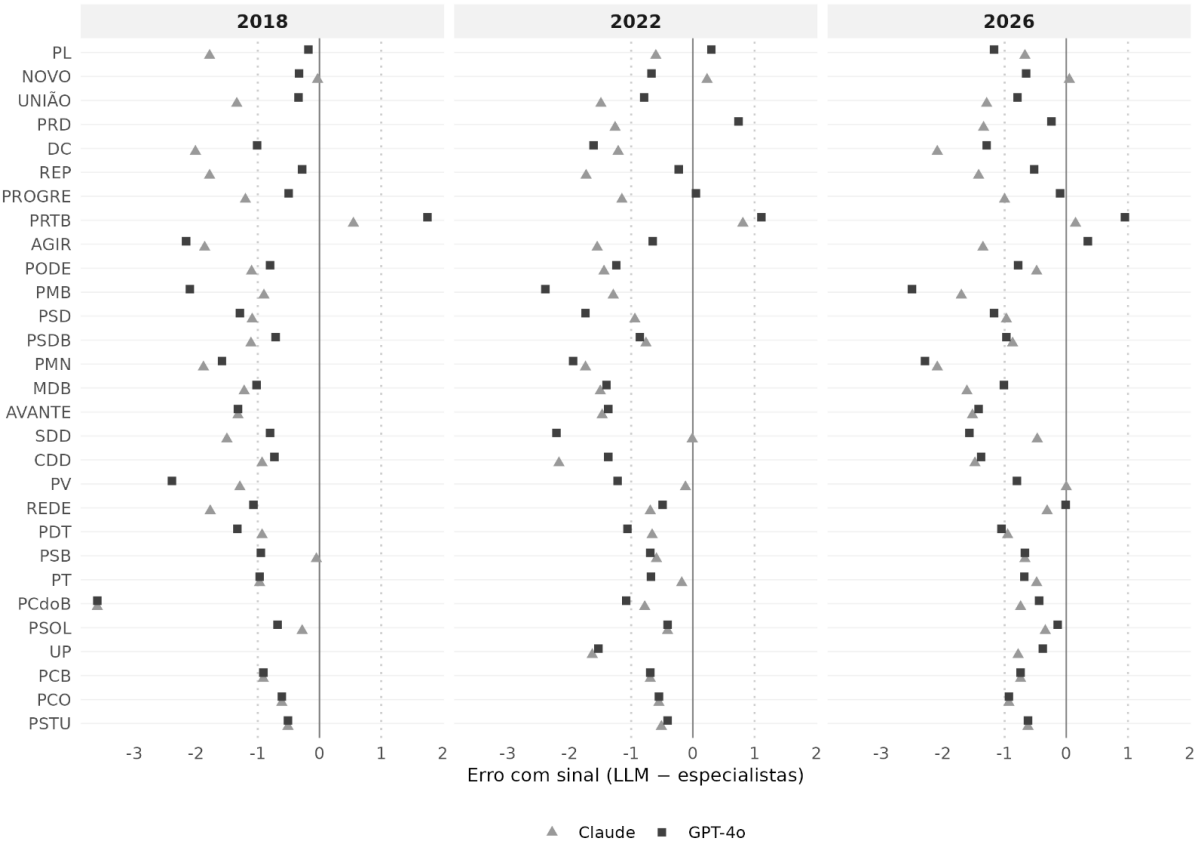


Nota. A linha tracejada indica concordância perfeita. A linha verde representa o ajuste linear. A faixa cinza destaca o intervalo central da escala, entre 4,5 e 5,5. O Missão foi excluído. *Fonte. elaboração própria a partir dos surveys de especialistas de Bolognesi, Ribeiro e Codato de 2018, 2022 e 2026 e das estimativas de GPT-4o e Claude Sonnet 4.6, coleta própria de 2026.*

O deslocamento não é homogêneo ao longo do espectro, como mostra a Figura 2. Na esquerda, o erro médio é de −0,61 no Claude e de −0,73 no GPT-4o, e parte dele pode decorrer de um artefato de escala. Os modelos atribuem 0 a partidos da extrema esquerda, como PSTU, PCB e PCO, enquanto a média de centenas de especialistas raramente alcança o limite inferior da escala. Na centro-direita, entre 5,5 e 7,5, o erro chega a −1,10 no Claude e a −1,13 no GPT-4o, e não há artefato equivalente que explique esse padrão. É nessa faixa que se concentram AVANTE, PMN, MDB, PSD, SDD, PMB e CDD, partidos frequentemente associados no debate público ao rótulo fisiológico. Na direita, os modelos divergem um pouco mais. O Claude mantém deslocamento mais intenso, com −1,17, enquanto o GPT-4o se aproxima mais dos especialistas, com −0,58. A única exceção consistente é o PRTB, que ambos os modelos situam à direita da média dos especialistas nas três ondas.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

Figura 2. Erro assinado por partido, onda e modelo



Nota. Valores negativos indicam que o modelo posiciona o partido mais à esquerda do que os especialistas. Os partidos estão ordenados pela média dos especialistas. As linhas pontilhadas marcam ±1 ponto. *Fonte. elaboração própria a partir dos surveys de especialistas de Bolognesi, Ribeiro e Codato de 2018, 2022 e 2026 e das estimativas de GPT-4o e Claude Sonnet 4.6, coleta própria de 2026.*

A consequência mais visível do deslocamento aparece no centro da escala. Nas ondas de 2022 e 2026, os especialistas não situam nenhum partido entre 4,5 e 5,5, o que é coerente com o esvaziamento do centro documentado por Bolognesi (2024). Os modelos, ao contrário, colocam nessa faixa partidos que os especialistas situam sistematicamente na centro-direita, como AVANTE, PMN, MDB, CDD, PMB, PSD e SDD. Em 2018, quando os especialistas ainda viam um centro ocupado por REDE, PV e CDD, os modelos já o preenchiam com outros partidos, como AVANTE, PMN, SDD e PMB. Onde os especialistas enxergam um sistema polarizado e sem centro, os modelos enxergam um sistema com centro composto por partidos de programaticidade de baixa saliência. O paralelo com a França é direto. O maior desvio encontrado por Bol e Bono na escala esquerda-direita foi o do Renaissance, partido que os especialistas situaram à direita e o ChatGPT no centro. Nos dois países, portanto, os modelos deslocam para o centro partidos que a comunidade especializada lê como de direita. Há ao menos duas interpretações possíveis para esse padrão. A primeira localiza o problema

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

no repertório de linguagem mobilizado pelos modelos, que podem associar o vocabulário corrente do fisiologismo à ideia de centro ideológico. A segunda sugere que parte da divergência pode refletir a leitura específica da própria comunidade de especialistas.

Como já sustentado na versão de 2022 da classificação (Bolognesi *et al.* 2025), para o caso brasileiro, as federações foram centrais para o deslocamento dos partidos do centro para as franjas. Por outro lado, essa explicação não contempla o caso de comparação. O que pode explicar o comportamento dos especialistas nos dois casos é a necessidade de gradação da régua ideológica ao diferenciar legendas, levando a posições mais radicais do que se espera (Mair, 2001). Os dados deste artigo não permitem arbitrar entre essas interpretações, mas mostram que o centro reconstruído pelos modelos é substantivamente distinto daquele observado no *benchmark*.

3.3 Quanto do erro pode ser reduzido por calibração?

A seção anterior mostrou que boa parte do erro dos modelos é como uma régua deslocada, precisa, mas fora do lugar. Se isso for verdade, bastaria conhecer a posição correta de alguns partidos de referência para reposicionar essa régua e corrigir o restante do sistema. É essa a lógica da calibração por âncoras, tomamos alguns partidos cuja posição já é bem estabelecida e usamos o erro dos modelos nesses casos conhecidos para corrigir a estimativa dos demais partidos.

Tabela 3. MAD antes e depois da calibração por âncoras com ajuste de deslocamento

Onda	Modelo	Original	Validação cruzada	3 âncoras (mediana)	Âncoras fixas	Survey anterior
2018	GPT-4o	1,11	0,61	0,69	0,65	NA
2018	Claude	1,20	0,57	0,63	0,69	NA
2022	GPT-4o	1,02	0,64	0,67	0,73	0,48
2022	Claude	0,97	0,57	0,60	0,62	0,48
2026	GPT-4o	0,88	0,52	0,57	0,60	0,21
2026	Claude	0,94	0,49	0,52	0,52	0,21

Fonte. elaboração própria. Âncoras fixas avaliadas nos 23 a 25 partidos restantes, cujo MAD original varia de 0,95 a 1,27. Survey anterior corresponde ao MAD da regra de persistência.

A calibração funciona. Usando essa correção, o MAD cai entre 37% e 53%, e a proporção de partidos a menos de um ponto do valor real sobe de 41 a 66% para 79 a 90%. O resultado se mantém mesmo em cenários mais realistas, com apenas quatro partidos de referência conhecidos (PT, PSOL, PDT e PROGRE), a redução do MAD ainda fica entre 33% e 46%. Esse ganho, porém, tem uma condição importante, ele só existe porque alguém já sabe, de outra fonte, onde esses partidos-âncora realmente estão. A calibração não ensina o modelo a acertar sozinho, ela usa conhecimento externo

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

para corrigir um erro que já sabíamos existir. Depois de calibrados, os modelos deixam de inventar um centro ideológico onde os especialistas não veem nenhum, evidência de que esse centro artificial era, em grande parte, produto do próprio deslocamento da régua, e não uma leitura genuína dos partidos de centro.

Ainda assim, mesmo corrigidos, os modelos calibrados não superam consistentemente um *survey* de especialistas recente. Em 2022, logo depois de um período de mudanças relevantes, o modelo calibrado empata estatisticamente com a simples repetição da onda anterior. Em 2026, depois de um período mais estável, a repetição do passado continua sendo a opção mais precisa. A calibração ajuda, mas não transforma o modelo em substituto do *survey*.

3.4 Saliência partidária (H2)

A hipótese é intuitiva, quanto menos um partido é reconhecido, mais difícil deve ser para um modelo de linguagem posicioná-lo corretamente, porque simplesmente existe menos informação disponível sobre ele. Para testar isso, dividimos os partidos em grupos de saliência, usando como medida a taxa de especialistas que preferiram não opinar sobre cada legenda. Um partido que muitos especialistas se recusam a classificar é, presumivelmente, um partido pouco conhecido, mesmo entre pesquisadores da área. Nos partidos de alta saliência, o erro do GPT-4o fica entre 0,62 e 0,83 ponto. Nos poucos partidos classificados como de baixa saliência, apenas dois ou três por onda, o erro sobe para entre 1,01 e 1,87, mais que o dobro. A diferença é estatisticamente significativa para o GPT-4o em 2018 e 2022, mas não em 2026, e nunca é significativa para o Claude, resultado que deve ser lido com cautela porque o grupo de baixa saliência é pequeno demais para sustentar conclusões fortes por si só.

Tabela 4. Preditores do desvio absoluto com MQO e erros-padrão agrupados por partido

	M1	**M2**	**M3**	**M4. interação**
Taxa de não resposta (NR%)	0,037*** (0,012)	0,038*** (0,014)	0,038*** (0,013)	0,050*** (0,016)
Cadeiras (padronizado)	NA	0,002 (0,059)	0,005 (0,058)	0,005 (0,058)
Claude (ref. GPT-4o)	NA	NA	0,014 (0,107)	0,015 (0,104)
NR% × Claude	NA	NA	NA	−0,022 (0,018)
Onda 2026 (ref. 2018)	NA	NA	−0,262* (0,143)	−0,262* (0,144)
N / R^2	170 / 0,120	166 / 0,121	166 / 0,150	166 / 0,161

*Fonte. elaboração própria. * p<0,10; *** p<0,01. Erros-padrão agrupados por partido em 29 clusters. Em M4, NR% está centrado e o coeficiente corresponde à associação estimada para o GPT-4o.*

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

O resultado central atravessa as quatro colunas sem enfraquecer, cada ponto percentual a mais de não resposta está associado a um aumento de aproximadamente 0,038 ponto no erro do modelo, associação que se confirma mesmo em uma análise estatística alternativa mais rigorosa. Já o tamanho do partido, medido pelas cadeiras na Câmara, não apresenta nenhuma associação detectável em nenhuma das colunas. Em outras palavras, o que prediz o erro não é o partido ser grande ou pequeno, é o partido ser conhecido ou desconhecido, mesmo dentro da própria comunidade de especialistas.

A quarta coluna responde a uma pergunta mais específica, será que a saliência importa mais para um modelo do que para o outro? Isoladamente, a inclinação estimada é de 0,050 para o GPT-4o e de 0,027 para o Claude, o que poderia sugerir que o GPT-4o é mais sensível à saliência. Mas quando testamos diretamente se essa diferença entre os dois números é real ou apenas variação esperada por acaso, a resposta é que não há evidência suficiente para afirmar isso. Os dois modelos parecem responder à saliência de forma parecida, e o padrão observado nos testes de grupo anteriores provavelmente reflete o tamanho pequeno da amostra de partidos pouco salientes, não uma diferença genuína entre os modelos. No conjunto, os resultados sustentam H2, os LLMs realmente erram mais nos partidos menos conhecidos, e esse efeito não depende do tamanho eleitoral do partido nem varia de forma clara entre os dois modelos testados.

3.5 Mudança entre ondas (H3)

Para os 27 partidos presentes nas três ondas, classificamos cada combinação partido-modelo conforme a direção da mudança. A tipologia distingue direção correta, direção errada, previsão estática e ausência de mudança no *benchmark*. Essa distinção evita que uma resposta simplesmente congelada seja tratada como erro direcional equivalente a uma mudança no sentido oposto.

Tabela 5. Variação entre ondas, correlação e categorias de direção

Período	Modelo	r Pearson (p)	Correta %	Estática %	Correta entre os que mudaram %
2018→2022	GPT-4o	0,41 (0,03)	55,6	14,8	68,2
2018→2022	Claude	0,35 (0,08)	25,9	59,3	70,0
2022→2026	GPT-4o	−0,24 (0,23)	55,6	14,8	65,2
2022→2026	Claude	0,17 (0,41)	22,2	55,6	50,0
2018→2026	GPT-4o	0,14 (0,49)	55,6	14,8	65,2
2018→2026	Claude	0,54 (0,004)	33,3	51,9	69,2

Fonte. elaboração própria. N=27. A coluna "Correta entre os que mudaram" exclui previsões estáticas do denominador.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

Até aqui, os resultados mostraram uma fotografia, o quanto os modelos acertam a posição dos partidos em um único momento. Esta análise pergunta algo diferente, quando um partido muda de posição de uma onda para outra, o modelo percebe essa mudança? A resposta encontrada é preocupante, mas de um jeito específico. A correlação entre a mudança real, registrada pelos especialistas, e a mudança captada pelos modelos é fraca. O motivo principal não é que os modelos entendam errado a direção da mudança política, é que, boa parte do tempo, eles simplesmente não percebem que algo mudou. Chamamos isso de previsão estática, o modelo repete exatamente a mesma resposta de uma onda para a outra, como se o tempo não tivesse passado, mesmo quando o *prompt* pede explicitamente a posição em um período diferente. O Claude faz isso em mais da metade dos partidos, entre 52% e 59%, enquanto o GPT-4o faz isso em apenas 15% dos casos, uma diferença grande demais para ser coincidência. E quando o Claude de fato atualiza sua resposta, ele acerta a direção da mudança em praticamente 7 de cada 10 casos, tão bem quanto o GPT-4o. Ou seja, o problema do Claude não é interpretar mal a política brasileira, é não perceber que o cenário mudou.

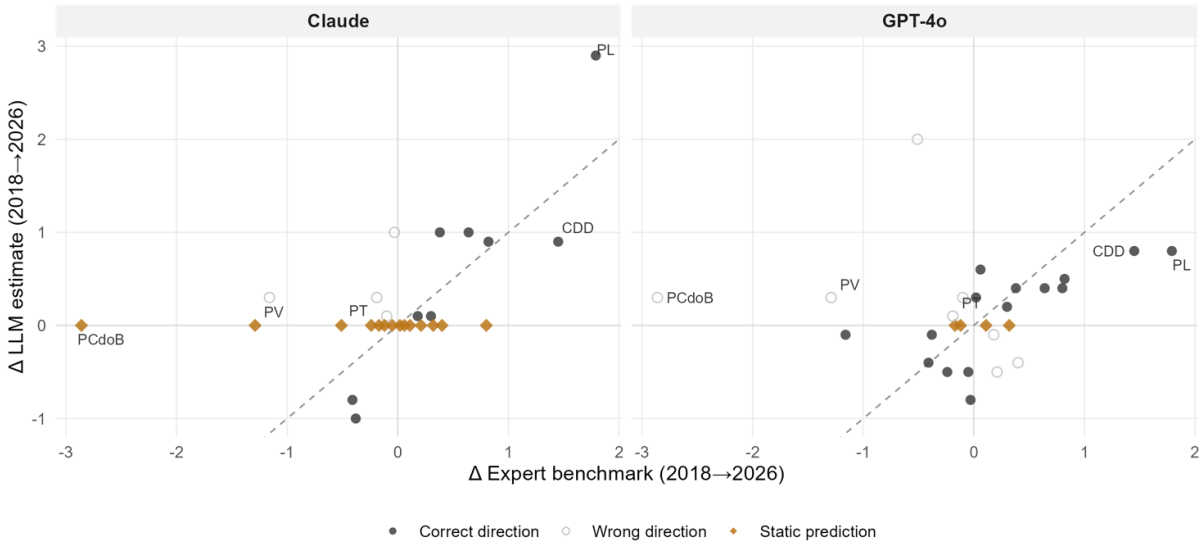
Os casos mais emblemáticos tornam esse padrão concreto. O PL passou de 7,78 para 9,57 entre 2018 e 2022, mudança que coincide com a filiação de Jair Bolsonaro ao partido, um evento amplamente noticiado. Os dois modelos captam esse deslocamento sem dificuldade. Já o PCdoB, cuja posição despenca de 4,60 para 1,74 depois da formação de uma federação com PT e PV, é tratado de forma muito diferente pelos dois modelos, o GPT-4o pelo menos tenta acompanhar a mudança, ainda que sem muita precisão, enquanto o Claude mantém o partido travado exatamente em 1,0 nas três ondas, como se a federação nunca tivesse acontecido. O PV, envolvido na mesma federação, repete esse padrão no Claude. A diferença entre os dois casos sugere uma explicação, embora este artigo não tenha casos suficientes para confirmá-la com segurança, mudanças que viram notícia, como a filiação de uma figura pública, parecem mais fáceis de um modelo de linguagem captar do que reorganizações institucionais, como uma federação partidária, cujo significado exige um tipo de conhecimento mais especializado sobre como o sistema político funciona.

A Figura 3 permite visualizar esse padrão partido a partido. Cada eixo não mostra a posição do partido, mas o quanto essa posição mudou entre 2018 e 2026, no *benchmark* e nas respostas do modelo. Se um modelo acompanhasse perfeitamente a mudança real, seus pontos cairiam sobre a linha tracejada. Os círculos preenchidos marcam os casos em que o modelo acertou a direção da mudança, os círculos vazios marcam os casos em que errou, e os losangos marcam as previsões estáticas, partidos para os quais o modelo respondeu a mesma coisa nas duas ondas. É fácil notar, sobretudo no painel

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

do Claude, uma fileira de losangos alinhados exatamente sobre o zero, o retrato visual de um modelo que, para boa parte dos partidos, simplesmente não se move.

Figura 3. Variação na média dos especialistas e na estimativa dos LLMs entre 2018 e 2026 por modelo



Nota. losangos indicam previsões estáticas; círculos cheios indicam direção correta; círculos vazios indicam direção errada. N=27. *Fonte. elaboração própria.*

3.6 Comparação entre os modelos (H4)

No agregado, GPT-4o e Claude apresentam desempenho muito semelhante. O teste de *Wilcoxon* pareado não identifica diferença no desvio absoluto, com p entre 0,55 e 0,82, nem no erro assinado, com p entre 0,22 e 0,90. As estimativas dos dois modelos apresentam correlação de 0,96 a 0,98, valor semelhante ao observado entre cada modelo e o *benchmark*. O fato de ambos apresentarem deslocamento na mesma direção é compatível com a hipótese de que o padrão não seja uma idiossincrasia de apenas uma arquitetura. O desenho, entretanto, não permite identificar a origem desse deslocamento.

A semelhança média esconde duas diferenças de comportamento. A primeira é a atualização temporal discutida na seção 3.5. A segunda é a dispersão entre iterações. O desvio-padrão médio é 0,08 no Claude, que devolve a mesma resposta em 80% das células, e 0,40 no GPT-4o, com repetição integral em 19%. Diferentemente do sugerido por Bol e Bono (2025), essa dispersão não funciona como indicador de qualidade no caso brasileiro, pois sua correlação com o desvio absoluto é próxima de zero nos dois modelos. H4, tratada como comparação exploratória, mostra equivalência média no erro e no viés assinado, acompanhada de diferenças relevantes na atualização temporal e na variabilidade das respostas.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

3.7 Síntese dos resultados por hipóteses

Tabela 6. Síntese dos resultados por expectativa

Expectativa	Resultado	Principal evidência
H1. Convergência	Sustentada na ordem; parcial na posição	ϱ de 0,95 a 0,96; MAD de 0,88 a 1,20 com deslocamento à esquerda; desempenho inferior à persistência; calibração reduz MAD em 37% a 53%
H2. Saliência	Consistente com H2 no conjunto dos modelos	NR% $\beta \approx 0{,}038$ com $p<0{,}01$; cadeiras sem associação detectável; interação NR%×modelo com $p=0{,}22$
H3. Mudança entre ondas	Parcialmente sustentada	Correlação fraca e instável; alguns casos salientes captados; Claude estático em 52% a 59% dos partidos
H4. Comparação entre modelos	Sem diferença média no erro; diferenças de comportamento	MAD e erro assinado semelhantes; correlação de 0,96 a 0,98; ancoragem temporal maior no Claude com McNemar $p \leq 0{,}005$

Fonte. elaboração própria com base nos resultados da seção 3.

CONSIDERAÇÕES FINAIS

Os resultados podem ser resumidos em uma formulação simples. Os LLMs reproduzem muito bem a ordenação do sistema partidário brasileiro, mas apresentam menor convergência na posição absoluta e desempenho limitado na atualização temporal. A ordenação é estável, a posição absoluta aparece deslocada para a esquerda em relação aos especialistas e parte substancial desse deslocamento pode ser reduzida quando há âncoras externas. A dimensão histórica é a mais problemática, sobretudo no Claude, que frequentemente trava na mesma estimativa entre ondas, mesmo quando o cenário muda.

Em relação a Bol e Bono (2025), o caso brasileiro reproduz alguns padrões e acrescenta novas restrições. O tamanho eleitoral não está associado a menor erro, e um dos desvios substantivamente mais importantes ocorre quando modelos aproximam do centro partidos que os especialistas situam à direita. O caso brasileiro acrescenta ainda três qualificações. A primeira é que a saliência relevante pode ser também a da própria legenda, e não apenas a do tema político. A segunda é que partidos posteriores ao corte de conhecimento configuram um problema de informação ausente, como mostra o Missão, e não um teste equivalente ao de legendas já presentes no repertório do modelo. A terceira é que a dispersão entre iterações não identifica, neste desenho, os partidos mais mal posicionados. O desenho em três ondas também oferece evidência compatível com a interpretação de Di Leo *et al.* (2025) de que os LLMs recuperam conhecimento político sedimentado. A forte repetição temporal do Claude é consistente com essa possibilidade, mas não identifica o mecanismo que a produz. O fato de 2018 ser a onda de maior erro também torna menos plausível uma explicação baseada apenas na

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

reprodução direta da classificação publicada naquele período, embora não elimine inteiramente essa possibilidade.

O deslocamento sistemático para a esquerda admite interpretações que este desenho não separa. Uma possibilidade é que as saídas dos modelos incorporem orientações políticas sistemáticas, fenômeno documentado em outros desenhos experimentais (Motoki *et al.*, 2024). Outra possibilidade é que modelos e especialistas utilizem referências substantivas diferentes para mapear o mesmo espaço ideológico. A divergência entre fontes deve, portanto, ser tratada como objeto de explicação, e não apenas como erro a ser minimizado. Os resultados também dialogam com três debates sobre o sistema partidário brasileiro. O primeiro envolve a distinção entre indefinição programática, centralidade governista e centrismo ideológico. O segundo diz respeito ao esvaziamento do centro observado nos *surveys*, que seria parcialmente revertido por uma aplicação não calibrada dos LLMs. O terceiro envolve a sensibilidade dos modelos a reorganizações institucionais, como as federações, que nos casos examinados não produziram atualizações consistentes.

As evidências sugerem que diferentes aplicações exigem níveis distintos de cautela. O Quadro 1 resume as condições de uso observadas neste desenho.

Quadro 1. Protocolo de uso de LLMs para posicionamento de partidos brasileiros

Aplicação	**Recomendação**
Ordenar partidos ou blocos no eixo esquerda-direita	Adequado para ordenação relativa neste desenho, com ρ de 0,95 a 0,96
Estimar posições, distâncias, polarização	Exige calibração externa quando a posição métrica é substantivamente importante
Posicionar partidos de baixa saliência	Usar com cautela. O erro aumenta com a não resposta entre especialistas
Partidos criados após o corte de treinamento	Não adequado sem informação externa, como ilustra o caso Missão
Substituir um survey recente	Não substitui a onda recente. A regra de persistência apresenta MAD menor em 2022 e 2026
Atualizar posições após mudanças	Promissor quando há mudança saliente e âncoras externas para calibração
Reconstruir posições históricas	Não recomendado sem validação histórica independente, sobretudo para Claude

Fonte. elaboração própria com base nos resultados da seção 3.

Para estimações correntes, os dois modelos apresentam desempenho agregado semelhante. Para reconstrução retrospectiva, o GPT-4o mostra menor ancoragem temporal entre os dois modelos testados, mas isso não o torna um substituto para dados históricos independentes. Combinar os modelos acrescenta pouco quando ambos compartilham o mesmo deslocamento. Mais importante do

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

que escolher entre eles é documentar a versão exata do modelo, a data de coleta, o *prompt* integral e os parâmetros de geração. Quando o objetivo envolve posições métricas, a calibração requer partidos-âncora definidos a partir de uma fonte externa e distribuídos ao longo do espectro. Também é recomendável permitir respostas de desconhecimento e evitar interpretar estabilidade entre iterações como evidência de validade. Como modelos comerciais são atualizados ou descontinuados, estudos reproduzíveis devem arquivar as respostas por iteração e, sempre que possível, incluir modelos de pesos abertos.

Quatro limitações orientam a interpretação. Primeiro, o *benchmark* utiliza médias agregadas e, sem os microdados, não é possível separar completamente o erro do modelo da incerteza das próprias avaliações dos especialistas. Segundo, o *proxy* de saliência baseada na não resposta foi medida em 2026 e aplicada retrospectivamente às três ondas, o que pressupõe uma estabilidade que precisa ser testada. Terceiro, o desenho compara apenas dois modelos comerciais e utiliza respostas em escala inteira. Quarto, não é possível descartar inteiramente a contaminação da onda de 2018 por classificações publicadas posteriormente, ainda que essa seja a onda de maior erro. Há ainda uma questão inferencial a verificar. Os modelos lineares usam erros-padrão agrupados em apenas 29 partidos, de modo que a versão final deve confirmar os resultados com uma correção apropriada para pequeno número de clusters ou com *bootstrap* por cluster.

A agenda de pesquisa inclui incorporar a dimensão de objetivos partidários associada ao fisiologismo, já medida pelos *surveys*, para testar se os modelos confundem indefinição programática com centrismo. Também é necessário comparar protocolos de *prompt*, ampliar o conjunto de modelos com pesos abertos e diferentes condições de acesso à informação, modelar a incerteza do *benchmark* com os microdados, repetir o teste depois da publicação integral da onda de 2026 e estender o desenho a outros sistemas partidários fragmentados da América Latina. A conclusão converge com a cautela proposta por Bol e Bono (2025). LLMs podem funcionar como fonte complementar de mensuração, mas sua validade depende de *benchmarks* independentes, e são justamente as classificações dos especialistas que permitem avaliar, calibrar e compreender as estimativas produzidas pelos modelos.

REFERÊNCIAS

BENOIT, K.; LAVER, M. Benchmarks for text analysis: a response to Budge and Pennings. *Electoral Studies*, v. 26, n. 1, p. 130-135, 2007.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

BENOIT, K.; DE MARCHI, S.; LAVER, C.; LAVER, M.; MA, J. Using large language models to analyze political texts through natural language understanding. *American Journal of Political Science*, publicado online em 2026. DOI 10.1111/ajps.70050.

BOL, D.; BONO, P.-H. Can ChatGPT accurately identify the position of parties? A validation study with an expert survey in France. *Research & Politics*, v. 12, n. 2, p. 1-9, 2025.

BOLOGNESI, B.; RIBEIRO, E.; CODATO, A.; SILVA, B. F. da. O desaparecimento do centro ideológico no sistema partidário brasileiro: a classificação mais atualizada dos experts. *Opinião Pública*, Campinas, v. 31, p. 1-28, e31120, 2025.

BOLOGNESI, B. Como medir ideologia partidária? *Revista de Sociologia e Política*, v. 32, e016, 2024.

BOLOGNESI, B.; RIBEIRO, E.; CODATO, A. Uma nova classificação ideológica dos partidos políticos brasileiros. *Dados*, v. 66, n. 2, 2023.

BUDGE, I. Expert judgements of party policy positions: uses and limitations in political research. *European Journal of Political Research*, v. 37, n. 1, p. 103-113, 2000.

BUDGE, I. *et al. Mapping policy preferences*: estimates for parties, electors, and governments 1945-1998. Oxford: Oxford University Press, 2001.

CASTLES, F. G.; MAIR, P. Left-right political scales: some "expert" judgments. *European Journal of Political Research*, v. 12, n. 1, p. 73-88, 1984.

DI LEO, R. *et al.* Mapping (a)ideology: a taxonomy of European parties using generative LLMs as zero-shot learners. *Political Analysis*, v. 33, n. 4, p. 456-463, 2025.

DORMUTH, I. *et al.* A cautionary tale about "neutrally" informative AI tools ahead of the 2025 federal elections in Germany. In: GUIDOTTI, R.; SCHMID, U.; LONGO, L. (ed.). *Explainable Artificial Intelligence*: xAI 2025. Cham: Springer, 2026. p. 64-85.

GRIMMER, J.; STEWART, B. M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, v. 21, n. 3, p. 267-297, 2013.

JOLLY, S. *et al.* Chapel Hill expert survey trend file, 1999-2019. *Electoral Studies*, v. 75, 102420, 2022.

LAVER, M.; BENOIT, K.; GARRY, J. Extracting policy positions from political texts using words as data. *American Political Science Review*, v. 97, n. 2, p. 311-331, 2003.

LE MENS, G.; GALLEGO, A. Positioning political texts with large language models by asking and averaging. *Political Analysis*, v. 33, n. 3, p. 274-282, 2025.

MAIR, P. Searching for the positions of political actors: a review of approaches and a critical evaluation of expert surveys. In: LAVER, M. (ed.). *Estimating the policy positions of political actors*. New York: Routledge, 2001.

MARKS, G. *et al.* Cross-validating data on party positioning on European integration. *Electoral Studies*, v. 26, n. 1, p. 23-38, 2007.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

MOTOKI, F.; PINHO NETO, V.; RODRIGUES, V. More human than human: measuring ChatGPT political bias. *Public Choice*, v. 198, n. 1-2, p. 3-23, 2024.

PETERS, B. Polistemics: evaluating LLMs as information mediators in politics & elections. *arXiv preprint*, arXiv:2607.25953, 2026.

POOLE, K. T.; ROSENTHAL, H. D-Nominate after 10 years. *Legislative Studies Quarterly*, v. 26, n. 1, p. 1-44, 2001.

POWER, T. J.; ZUCCO JR., C. Estimating ideology of Brazilian legislative parties, 1990-2005. *Latin American Research Review*, v. 44, n. 1, p. 218-246, 2009.

SLAPIN, J. B.; PROKSCH, S.-O. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, v. 52, n. 3, p. 705-722, 2008.

TAROUCO, G.; MADEIRA, R. M. Partidos, programas e o debate sobre esquerda e direita no Brasil. *Revista de Sociologia e Política*, v. 21, n. 45, p. 149-165, 2013.

ZUCCO JR., C. Ideology or what? Legislative behavior in multiparty presidential settings. *The Journal of Politics*, v. 71, n. 3, 2009.

ZUCCO JR., C.; POWER, T. J. Fragmentation without cleavages? Endogenous fractionalization in the Brazilian party system. *Comparative Politics*, v. 53, n. 3, p. 477-500, 2021.

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS DA PESQUISA: Os dados não estão publicamente disponíveis devido a restrições, mas podem ser obtidos mediante solicitação justificada ao autor correspondente.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL: Os autores utilizaram a ferramenta de inteligência artificial Claude Sonnet 5.5 (Anthropic) como apoio à revisão de códigos de programação em linguagem Python e à revisão técnico-gramatical do texto. Os modelos GPT-4o e Claude Sonnet 4.6, que constituem o objeto de análise do estudo, são descritos na seção 2.2 e não se confundem com o uso de IA como ferramenta de apoio.

FINANCIAMENTO: O presente trabalho contou com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná (Fundação Araucária) e do Instituto Nacional de Ciência, Tecnologia e Inovação em Representação e Legitimidade Democrática (INCT ReDem).

CONTRIBUIÇÃO DAS/DOS AUTORES/AS:

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.18923

Nilton Sainz: Conceptualization, Methodology, Software, Data curation, Investigation, Visualization, Writing-Original draft preparation, Writing-Reviewing and Editing.

Bruno Bolognesi: Conceptualization, Data curation, Investigation, Writing-Original draft preparation, Resources, Supervision, Validation, Writing-Reviewing and Editing.

DECLARAÇÃO DE CONFLITO DE INTERESSE: Os autores declaram não haver conflitos de interesse.

MINIBIOGRAFIAS DOS/DAS AUTORAS DO PAPER:

Nilton Sainz. Professor do Departamento de Ciência Política e do Programa de Pós-Graduação em Ciência Política da UFPR. Doutor em Ciência Política pela UFPR e pós-doutorando no DCP/FFLCH-USP. Pesquisador associado ao INCT ReDem. Coordenador do Portal da Classe Política e do VoxDem.

Bruno Bolognesi. Professor do Departamento de Ciência Política e do Programa de Pós-Graduação em Ciência Política da UFPR. Doutor em Ciência Política pela UFSCar e pesquisador associado ao INCT ReDem. Coordenador do Laboratório de Partidos e Sistemas (LAPeS).

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.