Inferência Textual *Zero-Shot* em Português: Especialização Linguística *vs.* Raciocínio Explícito em Modelos de Linguagem Compactos

Gabriel Silvestre Mancini[1] **, Marcos Lopes**[1,2]

[1]Programa de Educação Continuada em Engenharia (PECE)
Escola Politécnica – Universidade de São Paulo (USP)

[2]Departamento de Linguística
FFLCH – Universidade de São Paulo (USP)

gabrielsmancini@gmail.com, marcoslopes@usp.br

Abstract. *This work investigates the inferential capabilities of small and medium-sized language models (up to 9B parameters) on Natural Language Inference (NLI) in Brazilian Portuguese under a strictly zero-shot regime. Grounded in the formal semantic definition of entailment as truth-condition preservation, a deductive relation in Peirce's tripartite classification [Peirce 1932], we compare Portuguese-specialized models (Sabiá, Bode, Gaia) with recent multilingual models that incorporate explicit reasoning and knowledge distillation (Qwen3, Gemma 2, Ministral 3) on ASSIN 2 and on a controlled synthetic dataset grounded in lexical inclusion (hyponymy/hyperonymy). Results show that explicit reasoning and distillation are more decisive than monolingual adaptation alone, and that high accuracy on ASSIN 2 may overestimate inferential competence by conflating inductive distributional similarity with deductive truth-condition preservation.*

Resumo. *Este trabalho investiga a capacidade inferencial de modelos de linguagem de pequeno e médio porte (até 9B de parâmetros) na tarefa de Inferência em Linguagem Natural (NLI) em português brasileiro, em regime estritamente* zero-shot*. Partindo da definição formal de acarretamento como preservação de condições de verdade, uma relação dedutiva na classificação peirceana [Peirce 1932], comparamos modelos especializados em português (Sabiá, Bode, Gaia) com modelos multilíngues equipados com mecanismos explícitos de raciocínio e destilação de conhecimento (Qwen3, Gemma 2, Ministral 3) no ASSIN 2 e em um* dataset *sintético controlado, fundamentado em acarretamento por inclusão lexical (hiponímia/hiperonímia). Os resultados indicam que raciocínio explícito e destilação são mais decisivos do que adaptação monolíngue isolada, e que acurácia elevada no ASSIN 2 pode superestimar a competência inferencial ao confundir similaridade distribucional indutiva com preservação dedutiva de condições de verdade.*

1. Introdução

Na semântica formal, o acarretamento é definido como uma relação de preservação de condições de verdade entre sentenças: S_1 acarreta S_2 quando, nos casos em que S_1 é verdadeira, S_2 também é. Essa noção lógica estrita, que na tipologia peirceana corresponde à

dedução, o único tipo de inferência que garantidamente preserva a verdade [Peirce 1932], distingue-se radicalmente de outros tipos de relação entre sentenças, como a similaridade semântica distribucional, aferida *indutivamente* por graus de semelhança léxico-contextual de padrões observados em dados de treinamento.

É precisamente essa dissociação entre relação dedutiva (acarretamento) e reconhecimento indutivo (similaridade) que torna o Reconhecimento de Inferência Textual (NLI) um desafio em aberto, especialmente para modelos compactos (até 9B de parâmetros) em regime estritamente *zero-shot*. Dois paradigmas concorrentes disputam esse espaço: modelos *especialistas* em português, via pré-treinamento contínuo ou ajuste fino instrucional (Sabiá [Pires et al. 2023], Bode [Garcia et al. 2024], Gaia [Camilo-Junior et al. 2025]), e modelos *generalistas* multilíngues com mecanismos explícitos de raciocínio e destilação de conhecimento (Qwen3 [Yang et al. 2025], Ministral 3 [Liu et al. 2026], Gemma [Gemma Team 2024]).

Nesse cenário, o presente trabalho busca contribuir em duas frentes. Primeiro, avalia comparativamente onze LLMs no *benchmark* ASSIN 2 [Real et al. 2020] em regime *zero-shot*. Segundo, introduz um *dataset* sintético controlado, fundamentado em acarretamento por inclusão lexical (hiponímia/hiperonímia) e alinhado à definição formal mencionada anteriormente, para mitigar efeitos de familiaridade prévia dos modelos com o *benchmark* e isolar a competência inferencial genuína.

2. Fundamentação e Trabalhos Relacionados

2.1. Inferência lógica: dedução, indução e abdução

Peirce classifica as inferências lógicas em três tipos com estruturas de raciocínio distintas [Peirce 1932, Peirce 1934]. A **dedução** parte de uma regra geral e de um caso para uma conclusão necessariamente verdadeira. É o único tipo em que a verdade das premissas *garante* a verdade da conclusão. A **indução** generaliza a partir de casos observados para uma regra probabilística: opera ampliativamente e está sujeita a revisão pela experiência. A **abdução** gera hipóteses explicativas para fatos surpreendentes, produzindo conclusões apenas prováveis.

Essa tipologia é diretamente relevante para a tarefa de NLI. O acarretamento semântico, relação que NLI pretende reconhecer, é por definição dedutivo: $S_1 \Rightarrow S_2$ significa que não existe situação em que S_1 seja verdadeira e S_2 falsa. Modelos de linguagem, porém, são treinados por estimativa de máxima verossimilhança, um processo fundamentalmente *indutivo*: aprendem regularidades estatísticas dos dados de treinamento e generalizam a partir delas. A questão empírica central deste trabalho é, portanto, se o treinamento indutivo, com ou sem especialização em português, é capaz de produzir comportamento dedutivo robusto na verificação de condições de verdade.

2.2. Acarretamento semântico e inclusão lexical

Na semântica formal, a operacionalização linguística mais direta do acarretamento é a *inclusão de sentidos*: se um predicado contém um hipônimo, a hipótese obtida pela substituição desse hipônimo por seu hiperônimo é necessariamente acarretada. Por exemplo, *O lateral-esquerdo sofreu a falta* acarreta *O jogador sofreu a falta* porque o conjunto de situações em que um lateral-esquerdo sofre uma falta está contido no conjunto de

situações em que um jogador sofre uma falta. Essa relação é estritamente semântica: independe de implicaturas conversacionais, pressuposições ou conhecimento enciclopédico extralinguístico.

A distinção entre acarretamento (relação lógica, dedutiva, baseada em condições de verdade) e similaridade semântica distribucional (grau de semelhança léxico-contextual, inferida indutivamente) é central para interpretar tanto o *design* dos *benchmarks* quanto o comportamento dos modelos avaliados.

2.3. Benchmarks de NLI em português

Os principais recursos para NLI em português são o SICK-BR [Real et al. 2018], adaptação do corpus SICK do SemEval-2014 para o português brasileiro, e o ASSIN 2 [Real et al. 2020]. O ASSIN 2 representa um avanço metodológico importante ao adotar classificação binária (acarretamento *vs.* não-acarretamento) com sentenças simples extraídas do SICK-BR, reduzindo complexidades sintáticas e discursivas que dificultavam a avaliação controlada do raciocínio semântico.

No espectro supervisionado, o BERTimbau [Souza et al. 2020] com *fine-tuning* no ASSIN 2 estabeleceu o estado da arte com acurácia próxima a 0,90. Em contraste, [da Silva et al. 2023] demonstraram que heurísticas simples, sobreposição lexical, detecção de negação e diferença de comprimento entre premissa e hipótese, atingem F1 $\approx$ 0,71, evidenciando vieses estruturais no *benchmark* que permitem desempenho competitivo sem inferência semântica profunda. Esse resultado levanta a hipótese de que parte do desempenho dos modelos supervisados decorre de exploração de regularidades do *dataset* e não necessariamente de modelagem explícita de inferência lógica. O intervalo entre o piso heurístico (0,71) e o teto supervisionado (0,90) define o espaço no qual posicionamos a avaliação dos LLMs *zero-shot*.

2.4. Modelos compactos: especialização *vs.* raciocínio

Duas vertentes recentes competem no campo dos LLMs compactos. A primeira aposta em **especialização linguística**: o Sabiá [Pires et al. 2023] aplica pré-treinamento contínuo de LLaMA sobre corpus em português (ClueWeb 22); o Bode [Garcia et al. 2024] usa ajuste fino com LoRA em instruções traduzidas (Alpaca); o Gaia [Camilo-Junior et al. 2025] combina pré-treinamento contínuo (13B tokens em português) com *weight merging* para restaurar capacidade instrucional.

A segunda aposta em **raciocínio explícito e destilação**: o Qwen3 [Yang et al. 2025] integra *thinking mode* com orçamento computacional adaptativo (*Chain-of-Thought*); o Ministral 3 [Liu et al. 2026] usa *Cascade Distillation* e pós-treinamento com GRPO em cadeias de pensamento; o Gemma 2 [Gemma Team 2024] substitui a previsão de tokens por destilação de conhecimento a partir de modelos maiores (27B), enriquecendo representações inferenciais.

3. Metodologia

3.1. Datasets

ASSIN 2. O ASSIN 2 [Real et al. 2020] foi carregado via Hugging Face e pré-processado para eliminação de pares duplicados exatos: pares com mesma concatenação

premissa——hipótese em mais de uma partição foram removidos para evitar inflação artificial de métricas. O conjunto resultante contém 9.304 instâncias (4.670 não-acarretamento, 4.634 acarretamento).

Dataset sintético. Como instrumento complementar de validação, foi construído um *dataset* sintético de 60 pares de sentenças, todos no mesmo domínio lexical (futebol), fundamentado exclusivamente na definição formal de acarretamento por inclusão hiponímica. Nos 30 casos de acarretamento, a hipótese resulta da substituição de um hipônimo da premissa por seu hiperônimo (por exemplo, *lateral-esquerdo* $\rightarrow$ *jogador*); nos 30 casos de não-acarretamento, a relação hipônimo-hiperônimo é violada ou invertida. Os pares foram sinteticamente gerados por GPT-5.2 a partir de *prompt* restritivo e submetidos a revisão manual sistemática e em pares, que verificou: (i) a correção gramatical das sentenças; (ii) a validade da relação hiponímica entre os termos substituídos; (iii) a ausência de outros mecanismos inferenciais além da inclusão lexical (como negação, pressuposição ou implicatura); e (iv) a adequação dos casos de não-acarretamento, assegurando que a hipótese não decorra da premissa por nenhum outro mecanismo semântico formal. O modelo usado na geração não integra o conjunto avaliado. A inclusão desse conjunto justifica-se pela possibilidade de que modelos avaliados tenham tido acesso ao ASSIN 2 durante o treinamento, produzindo vieses de familiaridade. Por instrução explícita, o dataset sintético isola o comportamento dos modelos em inferências baseadas em inclusão lexical, sem regularidades superficiais do *benchmark*.

3.2. Avaliação

A avaliação é estritamente *zero-shot*: nenhum modelo foi ajustado, adaptado ou exposto a exemplos da tarefa. Cada par *premissa–hipótese* é incorporado a um *prompt* fixo que solicita resposta binária (`0` ou `1`). Os parâmetros de geração são determinísticos (`temperature=0`, `do_sample=false`, `max_new_tokens=4`). A extração da predição usa a expressão regular `^[^a-z]*([01])`, capturando o primeiro dígito binário e tolerando caracteres de formatação iniciais. Respostas sem correspondência são contabilizadas como erros de conformidade.

A consistência intra-modelo foi avaliada em subconjunto fixo de 500 pares do ASSIN 2, reexecutados cinco vezes sob condições idênticas a fim de verificar a estabilidade das decisões e distinguir flutuações estocásticas de comportamento inferencial sistemático.

3.3. Modelos avaliados

A Tabela 1 apresenta os 11 modelos avaliados, organizados em três paradigmas. A execução foi realizada no ambiente Google Colab com GPUs NVIDIA L4 e A100, em precisão BF16 (FP8 para variantes *Instruct* do Ministral 3).

4. Resultados e Discussão

4.1. ASSIN 2: desempenho global

A Figura 1 apresenta a acurácia global dos modelos no ASSIN 2. A linha tracejada indica o piso heurístico (0,70) estabelecido por [da Silva et al. 2023]; a linha pontilhada, o teto supervisionado do BERTimbau-large (0,90), estado da arte na tarefa [Souza et al. 2020].

Tabela 1. Modelos avaliados, agrupados por paradigma.

Modelo	Tam.	Cobertura	Paradigma
Sabiá-7B	7B	PT-BR	Pré-treinamento contínuo
Bode-7B	7B	PT-BR	*Fine-tuning* (LoRA)
Gemma-3-Gaia-PT-BR-4B	4B	PT-BR	Pré-treino + *weight merging*
Ministral-3B-Instruct (FP8)	3B	Multilíngue	SFT + ODPO
Ministral-3B-Reasoning	3B	Multilíngue	GRPO em CoT
Ministral-8B-Instruct (FP8)	8B	Multilíngue	SFT + ODPO
Ministral-8B-Reasoning	8B	Multilíngue	GRPO em CoT
Qwen3-4B-Instruct	4B	Multilíngue	Híbrido *instruct*
Qwen3-4B-Thinking	4B	Multilíngue	Híbrido *thinking*
Gemma-2-2B-IT	2B	Multilíngue	Destilação de conhecimento
Gemma-2-9B-IT	9B	Multilíngue	Destilação de conhecimento

Modelos multilíngues recentes lideram: o Qwen3-4B-Instruct atinge 90,8%, equiparando-se ao teto supervisionado em regime *zero-shot*. O Gemma-2-9B e variantes *Reasoning* do Ministral também superam 87%. Modelos especializados em português exibem comportamento heterogêneo: o Gaia alcança 84,4%, mas Sabiá-7B e Bode-7B sofrem colapso de classe (acurácia global $\approx 50\%$, resultado sem direcionamento inferencial).

4.2. ASSIN 2: acurácia por classe

A Tabela 2 desagrega as acurácias por tipo de julgamento. Observa-se uma assimetria recorrente: a maioria dos modelos erra mais na classe *Não Acarretamento*, tendendo a aceitar hipóteses que não são formalmente acarretadas pela premissa. Sabiá-7B e Ministral-8B-Instruct colapsam para a classe negativa (0,0% e 0,7% em acarretamento); Bode-7B colapsa para a positiva (99,9% em acarretamento, 7,2% em não-acarretamento). A diferença de desempenho nas classes *Acarretamento* e *Não-Acarretamento* é significativa ($p < 0,0001$) para todos os modelos de acordo com o Teste Qui-Quadrado de Independência com Correção de Yates, com $\alpha = 0,05$.

4.3. ASSIN 2: acurácia por faixa de similaridade semântica

A Tabela 3 apresenta a acurácia segmentada pelas faixas de similaridade semântica (escala STS do ASSIN 2, de 1 a 5). Conforme definido por [Real et al. 2020], o valor 1 representa sentenças "totalmente diferentes" e o valor 5, "significado virtualmente idêntico". Nas faixas 1–2 (sentenças totalmente distintas), os melhores modelos atingem acurácia próxima de 100%, pois a decisão de não-acarretamento é trivial: a dissimilaridade total torna a relação dedutiva evidentemente falsa. O desafio intensifica-se nas faixas 3–4 e 4–5, onde a alta sobreposição lexical pode induzir modelos dependentes de similaridade a classificar incorretamente pares de alta semelhança como acarretamentos, mesmo quando a relação lógica não se sustenta.

O fenômeno mais crítico ocorre na faixa 5 (paráfrases): Sabiá-7B (14,3%) e Ministral-8B-Instruct (13,4%) colapsam justamente quando as sentenças são virtualmente idênticas, revelando viés sistemático pela classe negativa. O Bode-7B exibe o padrão in-

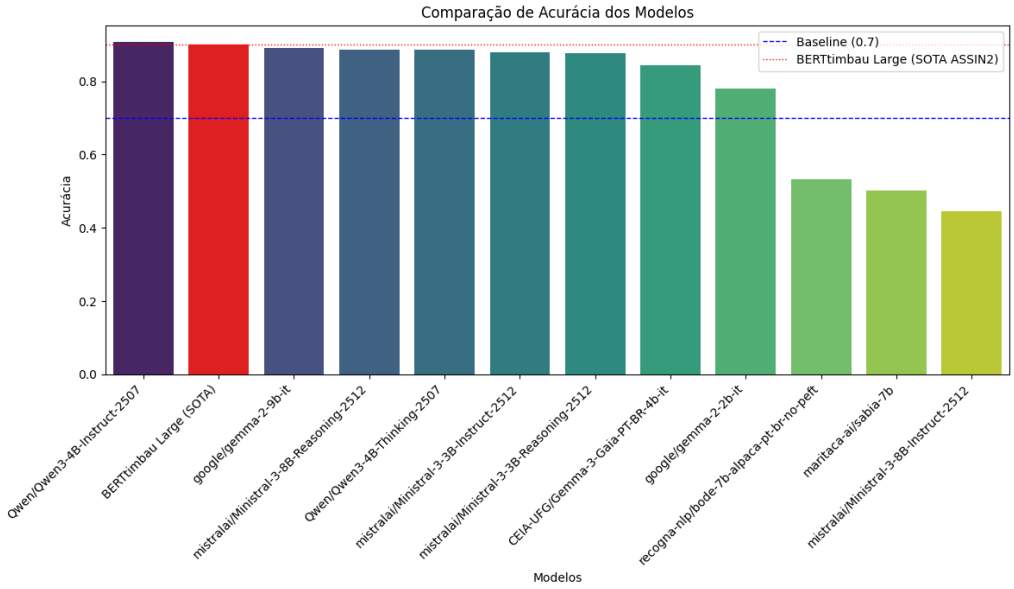

Figura 1. Acurácia global no ASSIN 2. Linha tracejada: piso heurístico (0,70); linha pontilhada: teto supervisionado BERTimbau-large (0,90).

verso: acerta bem na faixa 5 (86,5%) mas colapsa nas faixas 1–2 (10,1%), associando diretamente sobreposição lexical a acarretamento.

4.4. Dataset sintético: robustez sob controle semântico

A Tabela 4 apresenta os resultados no dataset sintético, construído exclusivamente com pares baseados em hiponímia. Nessa tabela, os valores de p referem-se à diferença de desempenho dos modelos nas classes *Acarretamento* e *Não Acarretamento*. Foram calculados aplicando-se o Teste Exato de Fisher (bicaudal), com $\alpha = 0{,}05$. O símbolo * indica significância estatística.

Qwen3-4B (*Instruct* e *Thinking*) e Gemma-2-9B mantêm acurácia $\geq$91,7%, demonstrando capacidade de generalização para a relação hiponímica sem dependência de regularidades do ASSIN 2. O Gaia também exibe robustez (90,0%), sugerindo que a especialização em português é mais eficaz quando combinada com uma base sólida de raciocínio. Em contraste, Sabiá, Bode e Ministral-8B-Instruct preservam o padrão de colapso, atingindo apenas 50%, acaso puro, em pares construídos para isolar a relação semântica formal.

4.5. Discussão

Padrões comportamentais e a distinção dedução/indução. A tipologia de Peirce oferece uma moldura interpretativa para os comportamentos observados. Modelos que colapsam para uma classe, Sabiá (sempre "não acarreta"), Bode (sempre "acarreta"), Ministral-8B-Instruct (sempre "não acarreta"), apresentam padrões de resposta incompatíveis com a verificação das condições de verdade que define o acarretamento: a decisão permanece constante independentemente do conteúdo proposicional do par. É importante ressaltar, contudo, que os experimentos medem exclusivamente comportamento observável em tarefas de classificação; inferir a partir dessas métricas a presença ou ausência de capacidades de raciocínio interno extrapola o que os dados permitem concluir diretamente. Esse

Tabela 2. Acurácia (%) por tipo de julgamento no ASSIN 2.

Modelo	Acarretamento	Não Acarret.	Global
Qwen3-4B-Instruct	94,7	86,8	**90,8**
Gemma-2-9B-IT	91,2	87,0	89,0
Ministral-8B-Reasoning	93,9	83,6	88,7
Qwen3-4B-Thinking	96,3	80,9	88,6
Ministral-3B-Instruct (FP8)	89,9	86,0	88,0
Ministral-3B-Reasoning	85,4	90,3	87,8
Gemma-3-Gaia-PT-BR-4B	95,9	72,9	84,4
Gemma-2-2B-IT	95,2	60,9	78,0
Bode-7B	99,9	7,2	53,4
Sabiá-7B	0,0	100,0	50,2
Ministral-8B-Instruct (FP8)	0,7	87,7	44,5

Tabela 3. Acurácia (%) por faixa de similaridade semântica (ASSIN 2).

Modelo	1–2	2–3	3–4	4–5	5	Global
Qwen3-4B-Instruct	98,5	93,3	92,4	88,3	91,3	90,8
Gemma-2-9B-IT	97,5	95,5	90,7	86,7	89,2	89,1
Ministral-8B-Reas.	94,9	91,7	89,3	85,7	89,9	88,7
Qwen3-4B-Thinking	99,5	96,2	91,2	83,5	89,9	88,6
Minist.-3B-Instr.(FP8)	95,5	95,7	90,8	85,5	87,8	88,0
Minist.-3B-Reas.	93,4	94,0	93,0	88,3	85,5	87,8
Gaia-PT-BR-4B	100,0	95,7	82,3	75,3	88,7	84,4
Gemma-2-2B-IT	98,5	82,6	69,3	65,9	86,3	78,0
Bode-7B	10,1	6,4	7,3	27,9	86,5	53,4
Sabiá-7B	100,0	99,8	97,6	78,9	14,3	50,2
Minist.-8B-Instr.(FP8)	95,5	94,3	86,5	67,3	13,4	44,5

comportamento reproduz as mesmas limitações das heurísticas simbólicas identificadas por [da Silva et al. 2023]: a distinção entre similaridade distribucional e preservação de condições de verdade permanece opaca para modelos sem mecanismos explícitos de raciocínio.

Raciocínio explícito e padrões de resposta mais robustos. Os melhores modelos, Qwen3, Gemma 2 e variantes *Reasoning* do Ministral, exibem padrões de resposta mais consistentes com os requisitos da tarefa: mantêm alto desempenho tanto nas faixas de baixa similaridade (rejeição correta de pares dissimilares) quanto na faixa 5 (aceitação correta de paráfrases), e preservam essa robustez nos dados sintéticos. Esses resultados sugerem que a *Chain-of-Thought* e a destilação de conhecimento produzem representações comportamentais mais alinhadas à distinção entre acarretamento e similaridade distribucional, embora não seja possível, a partir dos dados de classificação disponíveis, determinar se esse alinhamento resulta de raciocínio dedutivo genuíno ou de padrões aprendidos durante o treinamento que mimetizam esse comportamento.

Tabela 4. Acurácia (%) por tipo de julgamento no *dataset* sintético.

Modelo	Acarretamento	Não Acarret.	Global	p
Qwen3-4B-Thinking	90,0	100,0	**95,0**	0,2373
Qwen3-4B-Instruct	90,0	100,0	**95,0**	0,2373
Gemma-2-9B-IT	83,3	100,0	91,7	0,0522
Gemma-3-Gaia-PT-BR-4B	90,0	90,0	90,0	1,0000
Ministral-8B-Reasoning	73,3	100,0	86,7	0,0046 *
Ministral-3B-Instruct (FP8)	53,3	100,0	76,7	<0,0001 *
Ministral-3B-Reasoning	50,0	100,0	75,0	<0,0001 *
Gemma-2-2B-IT	96,7	40,0	68,3	<0,0001 *
Bode-7B	100,0	0,0	50,0	0 *
Ministral-8B-Instruct (FP8)	0,0	100,0	50,0	0 *
Sabiá-7B	0,0	100,0	50,0	0 *

A comparação entre Qwen3-4B-Thinking (88,6% no ASSIN 2 e 95,0% no sintético) e Qwen3-4B-Instruct (90,8% e 95,0%) indica que o *thinking mode*, embora marginalmente inferior no *benchmark*, é igualmente robusto em dados controlados. A superioridade do modo *instruct* no ASSIN 2 é estatisticamente significativa tanto no desempenho das classes *Acarretamento* e *Não-Acarretamento* isoladas quanto na média das duas ($p < 0,0001$ em Testes Qui-Quadrado de Homogeneidade, α = 0,05). Tal resultado parece sugerir uma exploração mais eficiente de padrões superficiais do *benchmark* por esse modelo. As diferenças se mostraram igualmente significativas nas comparações de mesmo tipo entre os modelos Gemma-2-9B-IT e Gemma-3-Gaia-PT-BR-4B. Entre eles, o ajuste fino em língua portuguesa do segundo modelo deve explicar os acertos na classe *Acarretamento*, negativamente compensada pelo mau desempenho na classe complementar. Considerando-se as duas classes, o modelo maior, sem ajuste fino específico para dados em português, acabou tendo resultados melhores globalmente.

A anomalia do Ministral-8B-Instruct. A disparidade entre Ministral-3B-Instruct (88,0%) e Ministral-8B-Instruct (44,5%) é contra-intuitiva perante as leis de escala. Com base nas informações do relatório técnico do Ministral 3 [Liu et al. 2026], uma hipótese explicativa é que o processo de alinhamento ODPO no modelo de 8B, ao otimizar a concisão e a segurança para assistentes de *chat*, possa ter suprimido padrões inferenciais mais elaborados, enquanto a destilação de *logits* específica aplicada ao modelo de 3B, combinada com a arquitetura de *tied embeddings*, teria mitigado esse efeito. É imprescindível, no entanto, ressaltar que essa interpretação é especulativa: não há experimentos de ablação ou avaliações controladas neste trabalho destinados a verificá-la empiricamente. Estudos futuros com variações sistemáticas dos parâmetros de alinhamento seriam necessários para confirmar ou refutar a hipótese.

Análise qualitativa de erros. A Tabela 5 ilustra como o colapso de classe se manifesta concretamente: Sabiá-7B nega o acarretamento mesmo quando a relação hiponímica é transparente (*lateral-esquerdo* $\subset$ *jogador*); Bode-7B afirma o acarretamento mesmo quando os predicados são de classes incompatíveis (*meia* $\not\to$ *árbitro*). O erro não de-

corre de dificuldades com o vocabulário, mas de viés decisório independente do conteúdo semântico do par.

Tabela 5. Predições corretas (✓) e incorretas (×) no dataset sintético.

Premissa	Hipótese	Acarret.	Sabiá	Bode
O lateral-esq. sofreu a falta	O jogador sofreu a falta	sim	(×)	(✓)
O meia organizou o meio-campo	O árbitro organizou o meio-campo	não	(✓)	(×)

***Zero-shot vs.* supervisionado.** O Qwen3-4B-Instruct (90,8%) iguala o BERTimbau-large ajustado supervisionadamente no ASSIN 2 (0,90), sem qualquer exemplo de treinamento. Isso demonstra que LLMs modernos compactos conseguem atingir o teto do estado da arte supervisionado, eliminando o custo de anotação manual. A diferença reside no custo de inferência: o BERTimbau-large opera com ∼335M parâmetros e pode rodar em hardware de entrada; o Qwen3-4B, mesmo quantizado, exige GPUs de pelo menos 8GB de VRAM. Para cenários de prototipagem rápida ou escassez de dados anotados, os LLMs *zero-shot* oferecem, portanto, uma alternativa viável ao ciclo custoso de anotação e treinamento supervisionado.

5. Conclusão

Este trabalho avaliou comparativamente 11 LLMs compactos em NLI em português brasileiro, em regime estritamente *zero-shot*, contrastando especialização linguística e raciocínio explícito. A introdução de um *dataset* sintético baseado em acarretamento por inclusão lexical (hiponímia/hiperonímia), permitiu separar competência inferencial genuína da exploração de regularidades superficiais do *benchmark*. Os resultados indicam que raciocínio explícito e destilação de conhecimento são mais decisivos do que adaptação monolíngue isolada para a tarefa. Do ponto de vista da semântica formal, modelos que exibem colapso de classe estão substituindo a verificação dedutiva de condições de verdade, a relação que define o acarretamento, pelo reconhecimento indutivo de similaridade distribucional. Acurácia elevada no ASSIN 2 pode, portanto, superestimar a competência inferencial quando não verificada em dados construídos sob restrição semântica estrita.

As principais limitações são: (i) o dataset sintético de tamanho reduzido e escopo temático único (60 pares, domínio do futebol, restrito à relação de hiponímia / hiperonímia), propõe-se mais como prova de conceito para avaliação controlada do que como avaliação de todos os tipos de acarretamento; (ii) os experimentos avaliam o comportamento observável em classificação binária e inferir a presença de raciocínio dedutivo interno a partir dessas métricas extrapola o que os dados permitem concluir diretamente e (iii) as interpretações causais sobre modelos específicos (notadamente o Ministral-8B-Instruct) são hipóteses baseadas em relatórios técnicos, não verificadas empiricamente neste estudo.

Como trabalhos futuros, vislumbra-se: a ampliação do dataset sintético para outras relações semânticas (negação, quantificação, pressuposição, predicação, operadores modais); a realização de estudos de ablação para investigar as causas do colapso em modelos específicos e a avaliação em regime *few-shot* para quantificar o ganho decorrente de exemplos explícitos.

Referências

Camilo-Junior, C. G., Oliveira, S. S. T., Pereira, L. A., Amadeus, M., Scotti, R., Fazzioni, D., Novais, A. M. A., and Jordão, S. A. A. (2025). GAIA: An open language model for Brazilian Portuguese. Hugging Face repository.

da Silva, F., Craveiro, G., da Silva, V., and Vanzin, V. (2023). Towards analysis on textual inference at ASSIN-2 dataset. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 229–234. SBC.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Cândido Júnior, A., Jodas, D. S., Afonso, L. C. S., Guilherme, I. R., Penteado, B. E., and Papa, J. P. (2024). Introducing Bode: A fine-tuned large language model for Portuguese prompt-based task. arXiv:2401.02909.

Gemma Team (2024). Gemma 2: Improving open language models at a practical size. arXiv:2408.00118.

Liu, A. H. et al. (2026). Ministral 3. arXiv:2601.08584.

Peirce, C. S. (1932). *Collected Papers of Charles Sanders Peirce, Volume II: Elements of Logic*. Harvard University Press, Cambridge, MA. Eds. Hartshorne, C.; Weiss, P.

Peirce, C. S. (1934). *Collected Papers of Charles Sanders Peirce, Volume V: Pragmatism and Pragmaticism*. Harvard University Press, Cambridge, MA. Eds. Hartshorne, C.; Weiss, P.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 226–240. Springer.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). The ASSIN 2 shared task: A quick overview. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 406–412. Springer.

Real, L., Rodrigues, A., Vieira e Silva, A., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I. C. S., Stanojević, M., Souza, R., and de Paiva, V. (2018). SICK-BR: A Portuguese corpus for inference. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 303–312. Springer.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. Springer.

Yang, A. et al. (2025). Qwen3 technical report. arXiv:2505.09388.