REGISTRO DE PROJETO

PLATAFORMA DA **DIVERSIDADE LINGUÍSTICA BRASILEIRA**: DADOS LINGUÍSTICOS PARA UMA IA BRASILEIRA

Raquel Meister Ko FREITAG
Departamento de Letras Vernáculas - Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brasil

Marcia dos Santos Machado VIEIRA
Departamento de Letras Vernáculas - Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Rio de Janeiro, Brasil

Juliana Bertucci BARBOSA
Departamento de Linguística e Língua Portuguesa - Universidade Federal do Triângulo Mineiro (UFTM)
Uberaba, Minas Gerais, Brasil

Miguel OLIVEIRA JR.
Faculdade de Letras - Universidade Federal de Alagoas (UFAL)
Maceió, Alagoas, Brasil

Cleber ATAÍDE
Departamento de Letras - Universidade Federal de Pernambuco (UFPE)
Recife, Pernambuco, Brasil

Alana de Santana CORREIA
Attenty Sistemas de Software (Attenty)
Campinas, São Paulo, Brasil

Amanda POST DA SILVEIRA
Instituto de Ciências Humanas e Letras - Universidade Federal do Jataí (UFJ)
Jataí, Goiás, Brasil

André Britto de CARVALHO
Departamento de Computação - Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brasil

EDITORES
- Cláudia Brescancini (PUCRS)

AVALIADORES
- Josane Oliveira (UEFS)
- Loremi Penkal (UNICENTRO)

SOBRE OS AUTORES
- Raquel Meister Ko Freitag
  Conceitualização; Administração do Projeto; Investigação; Escrita – Rascunho Original; Escrita – Análise e Edição.
- Marcia dos Santos Machado Vieira
  Conceitualização; Administração do Projeto; Investigação; Escrita – Rascunho Original; Escrita – Análise e Edição.
- Juliana Bertucci Barbosa
  Conceitualização; Administração do Projeto; Investigação; Escrita – Rascunho Original; Escrita – Análise e Edição.
- Miguel Oliveira Jr.
  Investigação.
- Cleber Ataíde
  Investigação.
- Alana de Santana Correia
  Investigação.
- Amanda Post da Silveira
  Investigação.
- André Britto de Carvalho
  Investigação.
- Andréia Silva Araujo
  Investigação.
- Brayna Conceição dos Santos Cardoso
  Investigação.
- Claudia Andrea Rost Snichelotto
  Investigação.
- Eduardo Cardoso Martins
  Investigação.
- Eliabe dos Santos Procópio
  Investigação.
- Elisa Battisti
  Investigação.
- Elisângela Nogueira Teixeira
  Investigação.
- Fabiane Cristina Altino
  Investigação.
- Hadinei Ribeiro Batista
  Investigação.
- Hendrik Teixeira Macedo
  Investigação.
- Isabel de Oliveira e Silva Monguilhott
  Investigação.

- Iury Cleveston
  Investigação.
- Kendra Dickinson
  Investigação.
- Lilian Cristine Hübner
  Investigação.
- Luma da Silva Miranda
  Investigação.
- Mailce Borges Mota
  Investigação.
- Marcus Garcia de Sene
  Investigação.
- Marinete Rodrigues da Silva
  Investigação.
- Marta Deysiane Alves Faria Sousa
  Investigação.
- Monica Maria Guimarães Savedra
  Investigação.
- Pedro Ricardo Bin
  Investigação.
- Ronice Muller de Quadros
  Investigação.
- Sandro Marcío Drumond Alves Marengo
  Investigação.
- Silvana Silva de Farias Araújo
  Investigação.
- Túlio Sousa de Gois
  Investigação.
- Valéria Viana Sousa
  Investigação.
- Valter de Carvalho Dias
  Investigação.

Recebido: 11/05/2025
Aceito: 14/09/2025
Publicado: 29/12/2025

COMO CITAR

FREITAG, R. M. K.; VIEIRA, M. S. M.; BARBOSA, J. B.; OLIVEIRA JR., M.; ATAÍDE, C.; CORREIA, A. S.; POST DA SILVEIRA, A.; CARVALHO, A. B.; ARAUJO, A. S.; CARDOSO, B. C. S.; SNICHELOTTO, C. A. R.; MARTINS, E. C.; PROCÓPIO, E. S.; BATtISTI, E.; TEIXEIRA, E. N.; ALTINO, F. C.; BATISTA, H. R.; MACEDO, H. T.; MONGUILHOTT, I. O. S.; CLEvESTON, I.; DICKINSON, K. V.; HÜBNER, L. C.; MIRANDA, L. S.; MOTA, M. B.; SENE, M. G.; SILVA, M. R.; SOUSA, M. D. A. F.; SAVEDRA, M. M. G.; BIN, P. R.; QUADROS, R. M.; MARENGO, S. M. D. A.; ARAÚJO, S. S. F.; GOIS, T. S.; SOUSA, V. V.; DIAS, V. C. (2025). Plataforma da Diversidade Linguística Brasileira: dados linguísticos para uma IA brasileira. *Cadernos de Linguística*, v. 6, n. 4, e863.

VERIFICAR ATUALIZAÇÕES

Andréia Silva ARAUJO
Departamento de Letras e Artes - Universidade Estadual de Santa Cruz (UESC)
Ilhéus, Bahia, Brasil

Brayna Conceição dos Santos CARDOSO
Faculdade de Letras - Universidade Federal do Pará (UFPA)
Abaetetuba, Pará, Brasil

Claudia Andrea Rost SNICHELOTTO
Programa de Pós-Graduação em Estudos Linguísticos - Universidade Federal da Fronteira Sul (UFFS)
Chapecó, Santa Catarina, Brasil

Eduardo Cardoso MARTINS
Faculdade de Letras - Universidade Federal do Amazonas (UFAM)
Manaus, Amazonas, Brasil

Eliabe dos Santos PROCÓPIO
Departamento de Letras Vernáculas - Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brasil

Elisa BATTISTI
Instituto de Letras - Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre, Rio Grande do Sul, Brasil

Elisângela Nogueira TEIXEIRA
Departamento de Letras Vernáculas - Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brasil

Fabiane Cristina ALTINO
Centro de Letras e Ciências Humanas - Universidade Estadual de Londrina (UEL)
Londrina, Paraná, Brasil

Hadinei Ribeiro BATISTA
Programa de Pós-Graduação em Linguística - Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Rio de Janeiro, Brasil

Hendrik Teixeira MACEDO
Departamento de Computação - Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brasil

Isabel de Oliveira e Silva MONGUILHOTT
Centro de Ciências da Educação - Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brasil

Iury CLEVESTON
Attenty Sistemas de Software (Attenty)
Campinas, São Paulo, Brasil

Kendra DICKINSON
Department of Spanish and Portuguese - Rutgers University (Rutgers)
New Brunswick, New Jersey, Estados Unidos

Lilian Cristine HÜBNER
Escola de Humanidades - Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre, Rio Grande do Sul, Brasil

Luma da Silva MIRANDA
Portuguese Departament - Eötvös Loránd University (ELTE)
Budapeste, Hungria, Brasil

Mailce Borges MOTA
Departamento de Língua e Literatura Estrangeiras - Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brasil

Marcus Garcia de SENE
Departamento de Linguística e Práticas de Ensino - Universidade de Pernambuco (UPE)
Garanhuns, Pernambuco, Brasil

Marinete Rodrigues da SILVA
Centro de Educação e Letras - Universidade Federal do Acre (UFAC)
Cruzeiro do Sul, Acre, Brasil

Marta Deysiane Alves Faria SOUSA
Curso de Licenciatura em Letras - Instituto Federal de Sergipe (IFS)
São Cristóvão, Sergipe, Brasil

Monica Maria Guimarães SAVEDRA
Instituto de Letras - Universidade Federal Fluminense (UFF)
Niterói, Rio de Janeiro, Brasil

Pedro Ricardo BIN
Programa de Pós-Graduação em Inglês - Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brasil

Ronice Muller de QUADROS
Centro de Comunicação e Expressão - Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brasil

Sandro Marcío Drumond Alves MARENGO
Departamento de Letras Vernáculas - Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brasil

Silvana Silva de Farias ARAÚJO
Departamento de Letras e Artes - Universidade Estadual de Feira de Santana (UEFS)
Feira de Santana, Bahia, Brasil

Túlio Sousa de GOIS
Departamento de Computação - Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brasil

Valéria Viana SOUSA
Departamento de Estudos Linguisticos e Literários - Universidade Estadual do Sudoeste da Bahia (UESB)
Vitória da Conquista, Bahia, Brasil

Valter de Carvalho DIAS
Curso de Licenciatura em Letras - Instituto Federal da Bahia (IFBA)
Salvador, Bahia, Brasil

RESUMO

A inteligência artificial generativa é baseada em modelos de língua em larga escala (LLMs), que são treinados com dados na maioria das vezes coletados sem consentimento ou infringindo direitos autorais. LLMs são treinados com bilhões de palavras e milhões de parâmetros, mas não sabemos exatamente quais textos são selecionados no treinamento ou quais parâmetros são controlados. Enquanto o aprendizado não supervisionado requer um grande volume de dados, demandando cada vez mais custos computacionais e gerando impactos energéticos, o aprendizado supervisionado, com dados estruturados e etiquetados pode otimizar esse processo; mais do que isso: o aprendizado

supervisionado com dados estruturados e etiquetados resultantes de projetos de documentação linguística podem contribuir diretamente para o Plano Nacional de Inteligência Artificial: "Desenvolver modelos avançados de linguagem em português, com dados nacionais que abarcam nossa diversidade cultural, social e linguística, para fortalecer a soberania em IA." No Brasil, além do português e suas variedades, há mais de 250 outras línguas (indígenas, de imigração, sinalizadas), negligenciadas na inclusão digital por falta de dados estruturados. O consórcio de laboratórios e grupos de pesquisa neste INCT visa a preparação de dados linguísticos para o treinamento de LLMs, considerando a diversidade linguística brasileira, com o desenvolvimento de um protocolo conjunto de coleta de dados linguísticos em campo, a ser replicado nos grupos e laboratórios longitudinalmente, assim como procedimentos de transcrição, alinhamento e etiquetagem de dados linguísticos para a constituição de um conjunto de dados que represente a diversidade linguística brasileira, e a realização de estudos sobre processamento linguístico da diversidade para o ajuste fino dos LLMs, contribuindo para a redução de assimetrias e preconceito resultantes do treino de LLMs com traduções do inglês.

PALAVRAS-CHAVE
LLM; Linguística; IA; Ciência de Dados.

TITLE
BRAZILIAN LINGUISTIC DIVERSITY PLATFORM: LINGUISTIC DATA FOR A BRAZILIAN AI

ABSTRACT
Generative artificial intelligence is based on large-scale language models (LLMs), which are trained with data most often collected without consent or in breach of copyright. LLMs are trained with billions of words and millions of parameters, but we don't know exactly which texts are selected in the training or which parameters are controlled. While unsupervised learning requires a large volume of data, demanding more and more computational costs and generating energy impacts, supervised learning with structured and tagged data can optimize this process; more than that: supervised learning with structured and tagged data resulting from language documentation projects can contribute directly to the National Artificial Intelligence Plan: "Develop advanced language models in Portuguese, with national data that encompasses our cultural, social and linguistic diversity, to strengthen sovereignty in AI." In

Brazil, in addition to Portuguese and its varieties, there are more than 250 other languages (indigenous, immigration, sign language), which are neglected in digital inclusion due to a lack of structured data. The consortium of laboratories and research groups in this INCT aims to prepare linguistic data for the training of LLMs, considering Brazil's linguistic diversity, with the development of a joint protocol for collecting linguistic data in the field, to be replicated in the groups and laboratories longitudinally, as well as transcription procedures, as well as procedures for transcribing, aligning and labeling linguistic data to create a data set that represents Brazilian linguistic diversity, and conducting studies on linguistic processing of diversity to fine-tune LLMs, helping to reduce asymmetries and prejudice resulting from training LLMs with translations from English.

KEYWORDS
LLM; Linguistics; AI; Data Science.

APRESENTAÇÃO[1]

A Plataforma da Diversidade Linguística Brasileira é uma iniciativa que visa documentar, preservar e disponibilizar o rico patrimônio linguístico do Brasil. O projeto surge como resposta à necessidade urgente de registrar e promover a diversidade de línguas e variedades linguísticas brasileiras, muitas das quais encontram-se em diferentes estágios de vitalidade e, em alguns casos, sob risco de desaparecimento.

O Brasil tem um cenário linguístico excepcionalmente diverso, que inclui não apenas o português brasileiro em suas múltiplas variantes regionais e sociais, mas também línguas indígenas, línguas de imigração, línguas de comunidades tradicionais e línguas de sinais. Em iniciativa reunindo pesquisadores, laboratórios e instituições, submetemos este projeto à chamada para INCT (Instituto Nacional de Ciência e Tecnologia) com o objetivo de construir um repositório digital abrangente e acessível, para pesquisadores, educadores, formuladores de políticas públicas e para toda a sociedade brasileira.

Linha de ações:

- **Mapeamento e identificação**: Realizar um levantamento abrangente das línguas e variedades linguísticas em situação de vitalidade no território brasileiro, estabelecendo prioridades para as ações de documentação linguística.
- **Padronização metodológica**: Elaborar um protocolo rigoroso para coleta, alinhamento, transcrição e etiquetagem de dados linguísticos que possa ser replicado pelos laboratórios e pesquisadores do INCT, bem como por colaboradores externos interessados em contribuir com o projeto.
- **Documentação linguística**: Conduzir ações sistemáticas de documentação para constituir um corpus representativo da diversidade linguística brasileira que será disponibilizado na plataforma.
- **Processamento de dados**: Preparar um extenso conjunto de dados linguísticos (15 milhões de tokens) com transcrição, alinhamento, anotação, etiquetagem e metadados adequados para disponibilização pública.
- **Análise de acessibilidade**: Testar os dados com medidas de processamento linguístico para avaliar sua acessibilidade, grau de intercompreensão e nível de dificuldade, considerando diversos perfis de usuários, com especial atenção a pessoas não escolarizadas, idosos e indivíduos com desenvolvimento atípico.

1 Projeto submetido à chamada CNPq/SECTICS/CAPES/FAPs Nº 46/2024 - Programa Institutos Nacionais de Ciência e Tecnologia – INCT. Uma versão de divulgação do *preprint* deste projeto, intitulada *Linguística para uma inteligência artificial (IA) brasileira*, foi publicada no *blog* SciELO em Perspectiva (https://blog.scielo.org/blog/2025/07/18/linguistica-para-uma-inteligencia-artificial-ia-brasileira/).

- **Desenvolvimento da Plataforma da Diversidade Linguística Brasileira**: Construir uma plataforma digital de acesso público e gratuito que disponibilize amostras linguísticas estruturadas e materiais suplementares destinados à educação linguística na sociedade brasileira.
- **Inovação tecnológica**: Treinar modelos de inteligência artificial para automatizar processos de transcrição e etiquetagem, permitindo a escalabilidade do conjunto de dados e a sustentabilidade do projeto no longo prazo.

A proposta foi aprovada no mérito mas não contemplada na faixa de recursos da chamada (pareceres de avaliação no APÊNDICE). Acreditamos que esta iniciativa representa um passo significativo de articulação de grupos de pesquisadores de diferentes áreas engajados em ações para a documentação científica, preservação e valorização da diversidade linguística brasileira, contribuindo para o reconhecimento deste patrimônio imaterial e para o desenvolvimento de políticas linguísticas mais inclusivas e representativas da diversidade cultural do Brasil.

1. TEMA ESTRATÉGICO

Inteligência Artificial. Além de Inteligência Artificial, a proposta do INCT tem aderência aos temas *Transformação Digital* e *Promoção da Igualdade e da Inclusão Social*, na medida que provê dados linguísticos estruturados contemplando a diversidade linguística do Brasil.

2. SETORES DE ATIVIDADE ECONÔMICA (CNAE) RELACIONADOS

- Tratamento de dados, provedores de serviços de aplicação e serviços de hospedagem na internet
- Ensino de idiomas
- Desenvolvimento e licenciamento de programas de computador customizáveis
- Atividades de bibliotecas e arquivos

3. OBJETIVO GERAL

A proposta deste INCT é desenvolver um protocolo padronizado de coleta de dados autênticos e multimodais que representem a diversidade linguística brasileira, abrangendo não só o português e suas variedades, mas também línguas de imigração, sinalizadas e indígenas, adotando práticas éticas e metodologicamente robustas para garantir a qualidade, autenticidade e representatividade dos

dados linguísticos coletados. Com as contribuições da sociolinguística para contemplar a variação linguística decorrente de fatores sociais, geográficos e culturais, e da psicolinguística quanto aos processos cognitivos envolvidos na produção, compreensão e aquisição da linguagem, o protocolo será estruturado para capturar dados de diferentes modalidades de comunicação (oral, escrita, sinalizada, gestual e visual) em contextos naturais, garantindo que os registros linguísticos reflitam o uso real das línguas em suas funções sociais. Após a coleta, os dados serão preparados, seguindo padrões de transcrição, etiquetagem e metadados organizados de acordo com os princípios FAIR, para a replicabilidade, interoperabilidade e uso compartilhado em pesquisas interdisciplinares. Os dados coletados e processados alimentarão a Plataforma da Diversidade Linguística, repositório que subsidiará estudos de descrição linguística, já que a disponibilização de um conjunto robusto e variado de dados permitirá não só o treino de LLMs, mas também descrições linguísticas mais precisas e com maior poder explanatório. A segurança na generalização dos resultados será reforçada pela diversidade e autenticidade dos dados, possibilitando avanços no entendimento das estruturas linguísticas, variação e mudanças no contexto brasileiro e do modo como a diversidade linguística é processada e representada. Os dados estruturados fornecerão uma base de treinamento de alta qualidade para o desenvolvimento de LLMs que reflitam e respeitem a diversidade linguística nacional, contribuindo para uma IA inclusiva.

4. OBJETIVOS ESPECÍFICOS

- Mapeamento de línguas em situação de vitalidade linguística nas bases legais e acadêmicas, para definir as línguas e variedades no Brasil a serem documentadas para a Plataforma da Diversidade Linguística Brasileira.
- Elaboração de protocolo de coleta, alinhamento, transcrição e etiquetagem de dados linguísticos, com detalhamento do roteiro de perguntas para a geração dos dados, rotinas de transcrição e etiquetagem e orientações FAIR.
- Treinamento de equipe para a coleta de dados, com curso in loco nos laboratórios e grupos associados, visando reprodutibilidade e padronização do conjunto de dados a ser construído.
- Preparação de conjunto de dados de treino de IA.
- Treino de IA com novos dados, com transcrição e anotação automáticas de gravações das línguas e variedades.
- Disponibilização da Plataforma da Diversidade Linguística Brasileira à comunidade.

5. MODELO DE GESTÃO E GOVERNANÇA PARA O INCT

As ações de planejamento para a Plataforma da Diversidade Linguística Brasileira vem sendo desenvolvidas já desde 2021, na articulação entre o GT de Sociolinguística da Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística (ANPOLL) e a comissão da área de Sociolinguística da Associação Brasileira de Linguística (ABRALIN) (Freitag *et al*., 2021; Machado-Vieira *et al*., 2021; Freitag 2022; Sousa; Freitag, 2024; dentre outros), o que permitiu a construção de uma rede de laboratórios envolvidos no empreendimento. A governança do INCT é estruturada em três níveis, o comitê gestor, o comitê técnico-científico e o núcleo operacional.[2]

As responsabilidades do comitê gestor são: definir a visão, missão e objetivos estratégicos do projeto; aprovar planos de ação, alocação de recursos e orçamento; estabelecer políticas de compartilhamento de dados e ética; e monitorar e avaliar os resultados do projeto; selecionar e supervisionar os comitês técnico-científicos. O conselho gestor reunir-se-á trimestralmente, com reuniões extraordinárias se necessário. Em sendo implementado o projeto, os comitês técnico-científicos para os objetivos específicos serão constituídos:

Comitê de Dados: Definir padrões para coleta, formatação e armazenamento de dados, com a ontologia e o protocolo de coleta; Comitê Linguístico: treinamento e supervisão para a qualidade linguística e a representatividade dos dados (coleta de campo, transcrição e etiquetagem);

Comitê de Ética e Inclusão: Elaborar e acompanhar as diretrizes éticas da coleta e disponibilização de dados;

Comitê Técnico de IA: Desenvolver e validar os modelos de IA.

O Núcleo Operacional é sediado na UFS, responsável pela execução diária das atividades. A equipe é constituída por um bolsista gerente de projeto, com a contratação de bolsistas para operar o datacenter, coletar e organizar os dados e implementar os modelos de IA. Os três comitês trabalharão nas etapas de planejamento, para identificar as línguas e variedades prioritárias, e construir protocolos de ação; na etapa de coleta de dados, para organizar expedições de campo e parcerias com comunidades locais, gravar dados orais, respeitando protocolos éticos, realizar transcrição, segmentação e anotação; na etapa de processamento e treinamento, para pré-processar os dados e treinar modelos de IA focados em reconhecimento, tradução e síntese de fala; na etapa de validação, para testar os modelos com linguistas e comunidades locais, e garantir

2 Na proposta submetida, a composição de comitê gestor apresentada foi: Raquel Meister Ko Freitag, Coordenadora geral, Marcia dos Santos Machado Vieira, vice-coordenadora, Juliana Bertucci Barbosa e Miguel Oliveira Jr.

representatividade e precisão, e ainda, na etapa de distribuição, para disponibilizar os resultados por meio da plataforma, e publicar relatórios e artigos científicos.

A avaliação do INCT se dará por meio de seminários de avaliação semestrais, em eventos organizados pelo comitê gestor para este fim, e nos seminários do INCT/CNPq, aos 24 e 48 meses de projeto. O comitê gestor é responsável pela formação de parcerias e prospecção de financiamento, com a adesão de novas universidades parceiras, bem como o monitoramento de Editais públicos (CNPq, FINEP, CAPES), para a continuidade do projeto, e a busca por parcerias com empresas de tecnologia (Google AI, IBM Watson, por exemplo). Líderes de laboratório que aderiram ao INCT serão responsáveis pela guarda dos equipamentos e bens designados pelo projeto, e serão responsáveis pela seleção e acompanhamento de bolsistas nas suas equipes. A produção científica derivada do INCT terá a contribuição registrada de acordo com a taxonomia CRediT, reconhecendo o papel desempenhado por cada uma das pessoas envolvidas em cada uma das atividades que geraram produtos acadêmico-científicos. O não cumprimento das diretrizes pelos laboratórios associados será avaliado pelo comitê gestor, que pode decidir pelo desligamento do laboratório, do mesmo modo que é papel do comitê gestor avaliar a adesão de novos laboratórios.

6. RESULTADOS CIENTÍFICOS E TECNOLÓGICOS JÁ OBTIDOS PELAS INSTITUIÇÕES QUE COMPÕEM O INCT, NA TEMÁTICA DA PROPOSTA

O campo da Sociolinguística brasileira é particularmente produtivo nas ações de documentação linguística e constituição de amostras. Projetos de grande porte como NURC, PEUL, VARSUL, etc (com cerca de 300 horas de áudio, com transcrição e anotação, o que equivale ~15 milhões de tokens) contribuíram para a formação de recursos humanos altamente qualificados (linguistas que constituem este INCT tiveram sua formação direta ou indireta vinculada a um destes projetos) e bases para a descrição do português brasileiro e suas aplicações (educacionais e tecnológicas). Estas ações de documentação envolvendo várias instituições, no entanto, foram realizadas nos últimos 50 anos, com financiamento (FINEP, CNPq, CAPES) e não tiveram continuidade. O resultado é que as amostras padronizadas em larga escala, com poder de generalização maior, referem-se a um estado de língua anterior, o que pode ser útil para descrições gramaticais, mas não são apropriadas para o treino de modelos de IA.

Estes projetos demandaram tempo médio de 10 anos para a constituição; com a tecnologia atual, e com o nível de capilaridade que propomos, esperamos reduzir este período para 1 ano, com treino para escalaridade do processo, e atingir o volume de aprendizado automático. Outro aspecto a considerar é que os projetos de documentação já existentes recobrem realidades linguísticas de

grandes centros urbanos e do litoral, via de regra no entorno das instituições onde estão sediados. Com a expansão da educação superior, novas universidades e novos núcleos de pesquisa, este INCT se destaca por construir uma rede predominantemente nas regiões nordeste e norte, o que não só releva a consolidação da expansão como insere as novas instituições em redes de pesquisa socialmente sensível e em uma temática relevante como a de dados linguísticos para a IA.

Do mesmo modo, as ações da psicolinguística no Brasil têm se voltado nos últimos anos para pesquisas não-WEIRD, ou seja, fora do entorno dos laboratórios das universidades já consolidadas. Novos laboratórios de psicolinguística nas instituições de expansão no nordeste e norte do país, incluídos nesta proposta de INCT junto com os já consolidados, têm potencial de ampliar a diversidade de compreensão do processamento da linguagem humana em diferentes regiões e perfis sociais brasileiros, e contribuir no ajuste fino de modelos de IA. Já vislumbrando a importância estratégica dos dados linguísticos, desde 2020, o GT de Sociolinguística da ANPOLL e a comissão da área de Sociolinguística da ABRALIN vêm trabalhando em ações para a construção da Plataforma da Diversidade Linguística Brasileira; a adesão do GT de Psicolinguística, que também vem se dedicando à pesquisa não-WEIRD, soma forças nesta proposta.

7. GRAU DE INOVAÇÃO E POTENCIAL DE IMPACTO DOS RESULTADOS SOB O PONTO DE VISTA CIENTÍFICO, TECNOLÓGICO, ECONÔMICO E SOCIOAMBIENTAL NO CONTEXTO NACIONAL E INTERNACIONAL

A inteligência artificial generativa, baseada em modelos de língua em larga escala, representa uma das mais avançadas inovações tecnológicas da atualidade, com potencial transformador em diversas áreas do conhecimento e da sociedade. Esses modelos, treinados com bilhões de palavras e milhões de parâmetros, enfrentam desafios éticos e técnicos significativos, incluindo a coleta de dados muitas vezes sem consentimento ou infringindo direitos autorais, além de consumir imensos recursos computacionais, com impactos ambientais consideráveis. Nesse cenário, o Brasil tem a oportunidade de liderar iniciativas que não apenas enfrentem esses desafios, mas também promovam avanços significativos em ciência, tecnologia, economia e inclusão social.

A proposta de desenvolver LLMs que representem a diversidade linguística brasileira, incluindo mais de 250 línguas além do português e suas variedades, destaca-se pela inovação ao integrar a riqueza cultural e linguística do país ao avanço da inteligência artificial. Essa abordagem vai além do tradicional uso de dados WEIRD (ocidentais, educados, industrializados, ricos e democráticos), criando um conjunto de dados inclusivo e representativo da diversidade linguística brasileira. A articulação de laboratórios e grupos de pesquisa em Sociolinguística e Psicolinguística fortalece a

interdisciplinaridade e fomenta a criação de um protocolo inédito para coleta, transcrição, alinhamento e etiquetagem de dados, aplicável em nível nacional e internacional.

Do ponto de vista científico, a iniciativa pode impulsionar os estudos de processamento de linguagem natural, contribuindo para o desenvolvimento de LLMs mais precisos e menos enviesados. Tecnologicamente, o impacto reside na criação de soluções inovadoras para inclusão digital, acessibilidade e comunicação, especialmente para comunidades historicamente marginalizadas. Economicamente, o treino de modelos de IA que compreendam a diversidade brasileira fortalece a soberania tecnológica do país, reduzindo a dependência de tecnologias estrangeiras.

Do ponto de vista ambiental, a busca por dados para treinar LLMs cada vez maiores é insustentável para o planeta. Muito tem sido discutido sobre os impactos ambientais negativos da inteligência artificial devido aos seus algoritmos de aprendizado de máquina que consomem muita energia e às emissões associadas a isso. Repercutiu muito fortemente a posição de Sam Altman, CEO da OpenAI (ChatGPT), que reconheceu não saber realmente como medir as necessidades energéticas desta tecnologia. Estima-se que a cada 100 tokens processados é necessário 0,5 litro de água para resfriamento de datacenters. A Google reportou em 2023 um aumento líquido no seu consumo de água de 17% a mais do que em 2022, enquanto a Microsoft reconheceu aumento de 34% em 2021. Uma das grandes demandas de energia decorre da baixa precisão do treinamento dos LLMs com conjuntos de dados não estruturados cada vez maiores. Há estudos que mostram que é possível reduzir a energia sem diminuir a precisão com a adoção de treinamento supervisionado, com dados anotados (Yokoyama *et al*., 2023). Relevar a importância de conjuntos de dados linguísticos constituídos com rigor científico e estruturados contribui para o fomento de práticas sustentáveis no uso de recursos computacionais, relevando a importância deste INCT para a sustentabilidade do planeta.

Ainda, a proposta do INCT promove justiça social ao garantir que comunidades indígenas, quilombolas, imigrantes e pessoas com deficiência sejam incluídas no cenário digital. Internacionalmente, o Brasil pode se posicionar como líder na criação de modelos linguísticos multiculturais e multilinguísticos, servindo de exemplo para outros países com rica diversidade cultural e linguística. Essa iniciativa não apenas enriquece a pesquisa em inteligência artificial, mas também cria uma base sólida para a redução de assimetrias globais, contribuindo para uma IA verdadeiramente inclusiva e ética.

8. PLANO DE DIVULGAÇÃO CIENTÍFICA

Para a disseminação dos resultados do projeto, em aderência aos princípios de Ciência Aberta, são estruturadas ações de divulgação em dois eixos, para especialistas e público amplo.

Para especialistas:

- Relatórios técnicos e resumos executivos para órgãos financiadores.
- Publicação de tutoriais, em aderência aos princípios de Ciência Aberta, com a publicação dos protocolos e ontologia em repositórios acadêmicos de acesso aberto, como Zenodo, ArXiv, em repositórios institucionais e no portal da Plataforma da Diversidade Linguística Brasileira. Um website do INCT será criado com o link para os repositórios, bem como para os guias e manuais de uso dos protocolos.
- Publicações acadêmicas, com divulgação sistemática dos resultados em periódicos de acesso aberto e com alto impacto em linguística e IA com creditação CRediT, em que a colaboração de cada participante é especificada na autoria dos produtos.
- Participação em conferências nacionais e internacionais (ABRALIN, ACL, LREC, SBC), com sessões específicas sobre o tema do INCT.
- Capacitação, com treinamentos para as equipes do INCT, em que serão disponibilizadas vagas para pessoas externas ao projeto.
- Acesso aos modelos, em seções de tutoriais e FAQs sobre o reuso dos dados no site do INCT, e em uma API aberta para pesquisadores e desenvolvedores utilizarem os dados.

Na divulgação para o público amplo, será criada a Plataforma da Diversidade Linguística Brasileira, um portal interativo com consulta pública aos conjuntos de dados e exemplos de uso em ensino de línguas e pesquisa, bem como materiais educativos explicando a diversidade linguística brasileira. São previstas ações de divulgação em mídias digitais, perfis do INCT no Blue Sky, Instagram, LinkedIn e YouTube para compartilhar resultados do projeto, convites para treinamentos e eventos, publicar vídeos curtos sobre a diversidade linguística brasileira e IA, e estabelecer parcerias com divulgadores de ciência para ampliar o alcance.

9. CONTEXTO METODOLÓGICO

Quanto maior a diversidade de línguas e variedades, mais dados são necessários para cobrir variações fonológicas, sintáticas e semânticas; porém dados anotados e consistentes podem reduzir a necessidade de grandes volumes, pois o modelo aprende mais eficientemente com dados estruturados. A Sociolinguística e a Psicolinguística têm experiência e tradição na obtenção de dados linguísticos com padrões científicos sólidos e bem estabelecidos. Entretanto, apesar disso, as coletas de dados são realizadas com foco em objetivos de pesquisa específicos, o que leva a uma

diversidade de estratificações socioculturais, que dificulta ou até mesmo inviabiliza a cumulatividade de amostras para estudos mais ampliados.

Assim, os passos metodológicos a serem desenvolvidos neste INCT envolvem inicialmente uma ação de mapeamento de amostras e de construção de ontologia específica para a realidade multilíngue e socioculturalmente diversa do Brasil, a fim de definir o protocolo a ser reproduzido pelos laboratórios e pesquisadores associados.

- **Criação de ontologia:** Uma ontologia estabelece estrutura formal e organizada para representar os conceitos, as relações e os atributos envolvidos no processamento da linguagem, com a definição padrão dos conceitos envolvidos, a fim de garantir consistência na anotação dos dados posteriormente, bem como permitir a integração de dados multimodais, com camadas de áudio, vídeo, transcrição e trilhas de anotação de informação linguística (anotação morfossintática e semântica) e incluir metadados organizados quanto ao contexto da amostra (quem falou, onde, quando, em que condições, geolocalização, etc).
  Uma ontologia é crucial para o trabalho reprodutível nos laboratórios das instituições envolvidas no INCT, bem como para servir como referência a outras iniciativas. Uma ontologia bem estruturada facilita o treinamento de algoritmos de aprendizado de máquina para tarefas como reconhecimento de fala, análise sintática e identificação de padrões linguísticos.

- **Mapeamento de línguas e variedades:** Para recobrir a realidade multilinguística e pluricêntrica do Brasil, em um cenário que estima a existência de cerca de 250 a 300 línguas, além da ontologia para a caracterização das amostras, é preciso estabelecer os critérios das línguas que serão consideradas na constituição das amostras. Do ponto de vista científico, os critérios sociolinguísticos priorizam as línguas em risco de extinção (número de falantes, perfil social, transmissão intergeracional), e pela situação de vitalidade e uso (contextos de uso, como família, comunidade e mídia, presença em instituições, como escolas e igrejas), a diversidade tipológica das línguas, localização geográfica e densidade populacional. No entanto, serão preponderantes os critérios de viabilidade e recursos, que consideram o acesso à comunidade e o consentimento, mas também a disponibilidade de financiamento. Cada laboratório prospectará línguas e variedades no seu raio de escopo, e os critérios serão incluídos nos metadados.

- **Protocolo de campo e preparação dos dados:** A fim de garantir a reprodutibilidade das ações nos laboratórios do INCT, bem como permitir a outros pesquisadores replicar ou expandir a amostra com base na estrutura estabelecida, será construído um protocolo de coleta de dados que considere a preparação do ambiente de gravação, rotinas éticas e legais, e o conjunto de instrumentos para conduzir as gravações, com itens e tarefas padronizados e reproduzíveis interlinguisticamente, para permitir comparações verticalizadas e aumentar a eficiência no treino de modelos.
  Para línguas e variedades pouco documentadas, como as que são alvo deste INCT, a etapa de transcrição, anotação e alinhamento, bem como revisão, é obrigatoriamente manual e realizada

por especialistas. Este processo demanda recursos humanos altamente especializados, recrutados e treinados nos laboratórios, o que demanda bolsas no financiamento.

Após transcritos, os dados serão anotados. O protocolo padronizará as camadas e as etiquetas. As amostras serão revisadas por transcritores e anotadores independentes, e validadas após concordância. Após a definição de protocolo, ações de treinamento de equipes de laboratórios devem ser realizadas para garantir a confiabilidade e padronização. As ações de coleta somente serão iniciadas após cada laboratório regularizar o projeto no CEP/CONEP. Para garantir a precisão nos dados, um kit de equipamentos de coleta (câmeras, microfones, placa integrada de som e computador), com as mesmas especificações, será direcionado para cada laboratório envolvido nesta etapa do projeto.

Após o procedimento de coleta e preparação, as amostras passarão à etapa de tokenização. O processamento é medido pela unidade *token*, que grosso modo corresponde a palavras. A meta do INCT é constituir uma amostra com entre 30?50 milhões de tokens, o que corresponde a 500-1000 horas de áudio, o que, após passar por transcrição e anotação, junto com metadados detalhados, é ponto de partida para treinar um modelo capaz de realizar transcrição e anotação no contexto da diversidade linguística brasileira.

- **Preparação dos dados para treino:** Após o pré-processamento, com a conversão e agrupamento de pares em entrada e saída, os dados serão utilizados para treinar um modelo para a automatização das próximas ações de transcrição e anotação de novas coletas, e constituir um conjunto de dados robusto para outras tarefas, ganhando escalabilidade. Para isso, é necessário configurar uma infraestrutura de processamento de dados, com a instalação de um servidor no DCOMP da Universidade Federal de Sergipe, para abrigar a Plataforma da Diversidade Linguística Brasileira.

  Se realizada hoje, a configuração partiria do modelo base de arquitetura Transformer, multitarefas, com camadas textual, áudio, vídeo e metadados para ajuste e configuração dos hiperparâmetros. O pareamento do modelo se dá por entradas de áudio/vídeo com saída de transcrição e transcrição com saída de anotação. O modelo será com métrica de Word Error Rate para transcrição e precisão/análise de concordância para anotação e rotinas de validação cruzada. Para o fine-tuning, serão utilizados dados de grupos específicos objetos de estudos psicolinguísticos, como pessoas idosas, com desenvolvimento atípico ou neurodivergência relacionada à leitura serão utilizados para o ajuste de modelo, assim como a ampliação dos metadados sociolinguísticos e de georreferenciação, para melhorar a acurácia. Nesta etapa, os resultados de pesquisas descritivas das línguas e variedades documentadas, decorrentes do desenvolvimento de pesquisas de pós-graduação, alimentarão a plataforma.

- **Ampliação da amostra e disponibilização na Plataforma da Diversidade Linguística:** Após o treino do modelo, no terceiro ano do INCT, a rotina de coleta será reiterada, transpondo as ações de preparação dos dados para o modelo treinado para transcrever e anotar novos dados, gerando escalabilidade para a disponibilização para uso em outras tarefas. Uma segunda aplicação do modelo treinado é o processamento das amostras linguísticas já constituídas em outros projetos, tais como NURC Digital, VARSUL, InCorpora, Falares Sergipanos, etc., com

possibilidade de inserção da dimensão temporal e exploração de modelos prospectivos para a deriva linguística, contribuindo para o ramo incipiente da Sociolinguística Computacional.

O conjunto de dados será disponibilizado à comunidade na Plataforma da Diversidade Linguística Brasileira, portal a ser constituído, com as amostras linguísticas disponibilizadas para o desenvolvimento de novas aplicações, mas também com o conjunto de materiais suplementares (protocolos de treinamento e ontologia) para subsidiar ações de descrição linguística, educação linguística e de revitalização, em especial daquelas línguas e variedades consideradas ameaçadas, assim como para uso pedagógico, subsidiando a implementação do direito de aprendizagem relacionado à diversidade linguística brasileira.

10. DISPONIBILIDADE E INFRAESTRUTURA

Este INCT envolve instituições fora dos grandes centros, em especial nas regiões Nordeste e Norte, e cujos laboratórios são coordenados por pesquisadores em sua maioria em fase de consolidação de carreira. O aporte de recursos para o estabelecimento de uma infraestrutura de pesquisa que seja padronizada e permita a reprodutibilidade é essencial para o sucesso da tarefa de constituição de amostras linguísticas. A infraestrutura do banco de dados VARSUL, por exemplo, foi constituída com financiamento da FINEP, entre 1989 e 1997, que envolvia a coleta de dados e o desenvolvimento de um sistema de armazenamento e compartilhamento (software interpretador/Engesis) (Freitag, 2021b).

Para a etapa de coleta e preparação dos dados linguísticos, cada laboratório associado ao INCT será equipado com um kit para gravação em campo, com um gravador de áudio portáteis de alta qualidade uma placa de interface de áudio USB para transmissão direta no computador, dois pares de microfones de lapela, dois pares de fones de ouvido de alta fidelidade para monitorar as gravações e suplementos de apoio (suportes para microfones, pop filters e cabos auxiliares XLR e TRS) e duas webcam HDMI. Para o processamento das gravações, cada laboratório será equipado com um computador com alta capacidade de processamento (CPU, RAM e SSD, a especificar no momento da compra, dadas os updates sistemáticos) para lidar com grandes volumes de dados, além de um dispositivo de armazenamento externo físico (servidor NAS) e backup na nuvem (AWS). Embora sejam em princípio equipamentos de baixo custo, para instituições menores é realmente difícil a aquisição destes equipamentos; e mesmo quando os laboratórios já tenham, as especificações são diversas, o que pode interferir na parametrização de uma ação de reprodutibilidade.

A infraestrutura para Treinamento de IA, a ser implementada na Universidade Federal de Sergipe, na sala de servidores do DCOMP/UFS, demanda a aquisição de um servidor com configuração avançada, apropriado para treinamento de redes neurais. Hoje, as configurações seriam processador AMD EPYC 9654P ou Intel Xeon W9-3495X (CPUs de alto desempenho com

muitos núcleos e threads, essenciais para suportar múltiplas GPUs e lidar com tarefas de pré-processamento), com placas de vídeo (GPUs) NVIDIA A100 (4 unidades) ou NVIDIA H100 (2 unidades), que são otimizadas para ML/LLM e oferecem performance superior para treinar redes neurais profundas com Tensor Cores e alta memória HBM, com memória RAM de 512 GB DDR5 ECC (Expansível para 1 TB), para garantir suporte para o treinamento e inferência em modelos com grandes conjuntos de dados, com armazenamento em SSD NVMe de 4 TB para alta velocidade de leitura/escrita e HDD de 20 TB para armazenamento de longo prazo.

Este será o maior investimento em infraestrutura, pois há demandas de pequenas adaptações estruturais no espaço físico, assim como a necessidade de prever recursos para a manutenção e reparo do equipamento, além de um gerador e nobreaks para garantir a continuidade do serviço em caso de queda de energia. Para o monitoramento do output, se faz necessária a aquisição de estações de trabalho ou servidores otimizados para aprendizado de máquina, também a serem instaladas na sala de servidores, para abrigar a Plataforma da Diversidade Linguística Brasileira. Esta plataforma, a ser configurada para ter diferentes níveis de acesso, desde o público amplo até pesquisadores em regime de associação e suas equipes, será o repositório para o armazenamento de novas amostras coletadas de acordo com o protocolo e seguindo a ontologia, a serem selecionadas a partir de critérios de curadoria, assim como o armazenamento de amostras constituídas previamente, para serem compartilhadas com usuários e para serem utilizadas no treino de modelo.

As etapas de pré-processamento de dados, com transcrição, normalização e limpeza de áudio, tokenização e alinhamento de texto com áudio, e registro de metadados, serão realizadas em laboratórios de informática multiusuários das instituições envolvidas. Os recursos de infraestrutura, no entanto, são investimento necessário, mas de menor valor neste INCT: o maior valor de investimento reside na formação de equipe de recursos humanos altamente especializados, envolvendo linguistas, para transcrição e anotação detalhadas, bem como pesquisadores das áreas da computação ligadas às ciência de dados, para o pré-processamento e treinamento de modelos de IA. O treinamento da equipe envolve não só a reprodução do protocolo de coleta e transcrição, mas também a formação para as boas práticas de gravação e etiquetagem, o uso de ferramentas para o processamento, assim como a formação para a conduta ética e legal, em respeito às diretrizes de ética em pesquisa e da lei geral de proteção aos dados, com consentimento livre e esclarecido, cuidados para a proteção de dados e atributos de licenciamento dos termos de uso do conjunto de dados (Creative Commons).

11. QUALIFICAÇÃO DO PROBLEMA SOB O PONTO DE VISTA CIENTÍFICO, TECNOLÓGICO E DE INOVAÇÃO

No documento do MCTI IA para o Bem de Todos, em que é apresentada a proposta de um Plano Brasileiro de Inteligência Artificial 2024-2028 (PBIA), um dos cinco objetivos listados é "Desenvolver modelos avançados de linguagem em português, com dados nacionais que abarcam nossa diversidade cultural, social e linguística, para fortalecer a soberania em IA." (MCTI 2024). Destacamos aqui a contribuição da Linguística para o pleno êxito deste objetivo. A Sociolinguística é o campo da ciência que estuda as relações entre língua e sociedade, e a Psicolinguística é o campo que estuda o processamento das línguas: o conjunto de trabalhos destes dois campos desenvolvidos no Brasil nos últimos 50 anos tem contribuições diretas para a consecução deste objetivo. Pensando em uma IA ética e socialmente sensível, a diversidade das comunidades na sociedade se reflete também (ou, pelo menos, deveria se refletir) na diversidade das comunidades em amostras linguísticas para treinar modelos de língua em larga escala.

Uma IA ética precisa atender aos princípios de justiça, equidade, diversidade e inclusão, e no domínio linguístico, por meio da seleção das amostras de línguas e variedades de línguas que vão compor o corpus de treino dos modelos, assimetrias se acentuam, desde a exclusão ou apagamento de línguas, até a priorização de uma variedade "dita de prestígio" face às variedades consideradas não-padrão ou estigmatizadas. Os preconceitos decorrentes dessa hierarquização de variedades são reproduzidos em LLMs e geram respostas (Shrawgi *et al.*, 2024; Fleisig *et al.*, 2024; Freitag; Gois, 2024), como já constatado no inglês afro-americano (Mengesha *et al.*, 2021).

Sobre língua, a Constituição de 1988 reconhece, no Art. 13., que "A língua portuguesa é o idioma oficial da República Federativa do Brasil." (Constituição Federal 1988). O objetivo do PBIA é ser o dispositivo legal. No entanto, não é apenas português que se fala no Brasil. A existência de outras línguas, embora empírica e legalmente reconhecidas, não faz parte do imaginário da nação, que se molda por uma ideologia monolíngue – a de que aqui todos falamos português – que se reproduz nos LLMs, na medida que somente o português é reconhecido como língua de soberania nacional no documento norteador.

Na própria Constituição, bem mais distante, há pistas da diversidade linguística, como no § 2º do Art. 210, que garante que "O ensino fundamental regular será ministrado em língua portuguesa, assegurada às comunidades indígenas também a utilização de suas línguas maternas e processos próprios de aprendizagem.", ou, ainda mais longe, no Art. 231., que diz que "São reconhecidos aos índios sua organização social, costumes, línguas, crenças e tradições, e os direitos originários sobre as terras que tradicionalmente ocupam, competindo à União demarcá-las, proteger e fazer respeitar todos os seus bens." Mesmo status de reconhecimento tem a Libras. O Art. 1º da Lei 10.436/2002

diz que "É reconhecida como meio legal de comunicação e expressão a Língua Brasileira de Sinais - Libras e outros recursos de expressão a ela associados." (Brasil 2002).

Além da informação de base legal sobre a existência de línguas, estudos linguísticos identificam e documentam outras tantas, de modo que não há consenso sobre quantas línguas são faladas no Brasil, nem quantas pessoas falam cada uma dessas línguas. Há, no entanto, consenso de que no Brasil não se fala apenas português, e uma política para a soberania nacional não deve ignorar a diversidade linguística, sob pena não só de excluir os povos originários, como também de excluir a identidade de uma população socialmente diversa.

LLMs para uma IA de soberania nacional precisam considerar a diversidade de línguas do Brasil, e não apenas eleger o português como língua de treino. E, mesmo dentro do português, há diversidade que reflete padrões sociais e culturais da realidade brasileira, que, como veremos na sequência, precisam ser considerados. Seja como uma das línguas com o maior número de falantes ou como uma língua com o maior número de países onde é falado, o português aparece nos ranqueamentos de línguas do mundo. O português não é apenas falado em Portugal e no Brasil. Não há um Português, há variedades de português, e cada uma destas variedades é polarizada em um centro, o que o configura o português como uma língua pluricêntrica.

O pluricentrismo do português é reconhecido nas ações de inclusão digital: é frequente encontrar documentação de software nas duas variedades hegemônicas do português (Português Europeu e Português Brasileiro). E, mesmo no Brasil, as especificidades de cada uma das comunidades que têm o Português como sua língua refletem seus valores socioculturais e diferenciam as variedades, o que tem sido amplamente demonstrado pela sociolinguística brasileira. A diversidade do português brasileiro é reconhecida no INDL "Língua Portuguesa e suas variações dialetais" e também é alçada a direito de aprendizagem na Base Nacional Comum Curricular Brasil (2018): "Compreender as línguas como fenômeno (geo)político, histórico, cultural, social, variável, heterogêneo e sensível aos contextos de uso, reconhecendo suas variedades e vivenciando-as como formas de expressões identitárias, pessoais e coletivas, bem como agindo no enfrentamento de preconceitos de qualquer natureza".

Para a soberania nacional, uma IA brasileira não pode se limitar a uma única língua, o português, nem a uma única variedade do português. O viés de seleção de uma única língua/variedade reforça e acentua ainda mais os preconceitos, em especial contra às variedades linguísticas subrepresentadas (Savedra; Mazzeli, 2020; Quadros *et al*., 2023; Meireles; Machado Vieira, 2023; Hamans *et al*., 2024; dentre outros).

O papel da linguística assenta-se na sua experiência em frentes de ações de documentação. Dados linguísticos autênticos, especialmente dados transcritos e anotados, têm aplicação em diversas áreas das tecnologias de linguagem e inteligência artificial. Atualmente estes acervos são armazenados de maneira assistemática e provisória, sem protocolos específicos para compartilhamento e reuso. Embora as instituições de pesquisa tenham repositórios para

compartilhamento de produção científica (teses, dissertações, etc.), estes não são apropriados para compartilhar coleções de dados linguísticos. A solução para este problema é a construção de um repositório próprio específico para este tipo de acervo, em uma iniciativa denominada Plataforma da Diversidade Linguística Brasileira, repositório nacional que oportunize à sociedade a diversidade do patrimônio linguístico que vem sendo registrado e mapeado. A plataforma visa a catalogação nacional (salvaguarda e difusão) e a constituição de um repositório comum, com padrões de metadados e diretrizes de armazenamento de coleções de dados linguísticos, de modo a atender necessidades tanto do público amplo, para que possibilite a qualquer interessado ver, ouvir, repetir diferentes manifestações de uso linguístico no país, como para público especializado, tal como a alimentação de LLMs para uma IA brasileira. Para isso, protocolos de coleta e preparação dos dados precisam ser desenvolvidos, bem como o treinamento de recursos humanos para reproduzi-los e alimentar de maneira sistemática a plataforma, a fim de subsidiar a disponibilização de dados longitudinais estruturado para o aprendizado supervisionado de máquina, em um perspectiva não-WEIRD, incluindo outros perfis que não pessoas ocidentais (Western), educadas (Educated), de países industrializados (Industrialized), ricos (Rich) e democráticos (Democratic).

Uma IA eticamente sensível para a soberania nacional requer que a diversidade linguística seja considerada de maneira plena, com amostras linguísticas diversificadas para o treino de LLMs. Sem isso, a reprodução de uma IA que considera apenas o português e uma de suas variedades tem efeito na conformação de padrões linguísticos hegemônicos, invisibilizando e marginalizando ainda mais as variedades linguísticas subrepresentadas.

12. INDICADORES/MARCOS

1º ano

1 Mapeamento e seleção das línguas e variedades a serem documentadas.
2 Protocolo de coleta de dados linguísticos multimodais.
3 Ontologia dos metadados da Plataforma da Diversidade Linguística Brasileira.

2º ano

4 Coleta de dados linguísticos em campo (transcrição e anotação manual).
5 Instalação do servidor para o treino de modelo.

3º ano

6 Preparação do dataset: Transcrição, alinhamento e etiquetagem dos dados linguísticos coletados em campo.
7 Disponibilização do dataset para treino, em repositório aberto.

4º ano

8 Coleta de dados linguísticos em campo (para transcrição e anotação automática)

5º ano

9 Ajuste fino do modelo (com dados psicolinguísticos)

10 Disponibilização da Plataforma da Diversidade Linguística Brasileira para acesso público.

INFORMAÇÕES COMPLEMENTARES

CONFLITO DE INTERESSE

Os autores declaram que não há conflito de interesse.

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS

Todo o conjunto de dados de apoio aos resultados deste estudo foi publicado no próprio artigo.

DECLARAÇÃO DE USO DE IA

Os autores declaram que nenhuma ferramenta de IA foi utilizada na escrita deste artigo.

AVALIAÇÃO E RESPOSTA DOS AUTORES

Avaliação: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID863.R
Resposta dos Autores: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID863.A

REFERÊNCIAS

FLEISIG, E.; SMITH, G.; BOSSI, M.; RUSTAGI, I.; YIN, X.; KLEIN, D. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. *In:* AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y-N (Eds.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* Miami, FL: Association for Computational Linguistics, 2024. p. 13541–13564.

FREITAG, R. Linguística para uma inteligência artificial (IA) brasileira. *SciELO em Perspectiva*, 2025. Disponível em: https://blog.scielo.org/blog/2025/07/18/linguistica-para-uma-inteligencia-artificial-ia-brasileira/. Acesso: 23 dez 2025

FREITAG, R. Gestão de acervos de documentação linguística: desafios, responsabilidades e planejamento. *In*: BRESCANCINI, C. (Org.). *Projeto Varsul – Variação Linguística do Sul do Brasil – 36 Anos.* Porto Alegre: Zouk Editora, 2021. p. 13–51, 2021.

FREITAG, R. Sociolinguistic repositories as asset: challenges and difficulties in Brazil. *The Electronic Library*, v. 40, n. 5, p. 607–622, 2022.

FREITAG, R. M. K.; MARTINS, M. A. R.; ARAÚJO, A.; BATTISTI, E.; COELHO, I. M. W. S.; SOUSA, M. D. A. F.; SILVA, R. G.; LIMA-LOPES, R. E. Desafios da gestão de dados linguísticos e a Ciência Aberta. *Cadernos de Linguística*, v. 2, n. 1, p. e307, 2021. DOI: 10.25189/2675-4916.2021.v2.n1.id307.

FREITAG, R. M. K.; DE GOIS, T. S. Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 15, 2024, Belém/PA. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024 . p. 317-326. DOI: https://doi.org/10.5753/stil.2024.241891.

HAMANS, C.; VAN DER VOORT, H.; OLIVEIRA JR., M.; VALDENILSON, K. The Economic Value of Endangered Languages. *Cadernos de Linguística*, v. 5, n. 1, e723, 2024.

MACHADO VIEIRA, M. S.; BARBOSA, J. B.; FREITAG, R. M. K.; BORGES, M. M.; MEDEIROS, A. L. S. Acervos de dados abertos à sociedade: memória linguística e sociocultural e potencialidade de (re)uso. *Cadernos de Linguística*, v. 2, n. 1, e607, 2021. DOI: https://doi.org/10.25189/2675-4916.2021.v2.n1.id607.

MEIRELES, V.; MACHADO VIEIRA, M. Variação em línguas românicas: ações do projeto VariaR como contributos de ciência aberta e cidadã. *Reflexos*, n. 6, p. 1-21, 2023. Disponível em: http://interfas.univ-tlse2.fr/reflexos/1325.

MENGESHA, Z.; HELDRETH, C.; LAHAV, M.; SUBLEWSKI, J.; TUENNERMAN, E. "I don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence*, v. 4, 725911, 2021.

QUADROS, R.; SILVA, J.; MACHADO, R. A corpus-based analysis of coordinate structures in Libras. *Advances in Sign Language Corpus Linguistics*, p. 108–123, 2023.

SAVEDRA, M.; MAZZELLI, L. Variedades linguísticas da imigração germânica no Brasil: vitalidade, glotopolítica e território. *A Cor das Letras*, v. 21, n. 1, p. 105–131, 2020.

SHRAWGI, H.; RATH, P.; SINGHAL, T.; DANDAPAT, S. Uncovering stereotypes in large language models: A task complexity-based approach. *In*: GRAHAM, Y.; PURVER, M. (Eds.). *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. St. Julian's, Malta: Association for Computational Linguistics, 2024. p. 1841–1857.

SOUSA, M.; FREITAG, R. Bancos de dados sociolinguísticos e a Ciência Aberta: compartilhamento de dados e conhecimentos. *Revista Diálogos*, v. 12, n. 1, p. 165–187, 2024.

APÊNDICE – Parecer CNPq

Resultado Final
Identificação da Proposta
Número do Processo:408788/2024-0
Solicitante: Raquel Meister Ko Freitag
Chamada:INCT_2024
Título do Projeto: Plataforma da Diversidade Linguística: Dados linguísticos para uma IA brasileira
Parecer de Deliberação final antes do período recursal

Critério:
Adequação da proposta, considerando: a qualificação do problema, a abordagem inter e transdisciplinar do tema enfocado, o grau de originalidade científica e tecnológica, a exequibilidade e a relevância para o desenvolvimento nacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 8.0

Critério:
Alinhamento com os temas estratégicos prioritários indicados nesta Chamada.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 2.0 **Nota:** 9.0

Critério:
Capacidade instalada das instituições integrantes do INCT para atuação em rede de pesquisa frente aos objetivos pretendidos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 8.0

Critério:
Caráter competitivo e mobilizador da proposta considerando as parcerias institucionais estabelecidas (apoio de agências de fomento, empresas e instituições do terceiro setor etc.).

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 7.0

Critério:
Capacidade e experiência do coordenador e da equipe de pesquisadores participantes em relação ao atingimento dos objetivos e metas propostos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 8.0

Critério:
Abrangência e sinergia das atividades e dos atores envolvidos na proposta, consideradas a complexidade dos temas abordados, a complementaridade de suas competências e a necessária abordagem multidisciplinar para a solução de problemas complexos, incluindo o setor empresarial e sociedade.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 7.0

Critério:
Adequação do orçamento e coerência do cronograma de execução em relação às metas e objetivos estabelecidos na proposta.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 6.0

Critério:
Estrutura operacional e modelo de gestão do INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 0.5 **Nota:** 8.5

Critério:
Adequação do Plano para Disseminação do conhecimento gerado pelo INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 0.5 **Nota:** 8.0

Critério:
Grau de inovação e potencial de impacto dos resultados esperados sob o ponto de vista científico, tecnológico, econômico e socioambiental no contexto nacional e internacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 7.5
Nota Final

Nota	Ordem	Prioridade
7,78		

Resultado da Avaliação
Desfavorável
Justificativa:
Com base na avaliação de mérito do Comitê Julgador, a proposta não atingiu prioridade suficiente para aprovação frente à demanda qualificada e ao orçamento disponível para a Chamada.
Data de Emissão
Data de Emissão do Parecer: 21/03/2025

Parecer de Recomendação

Critério:
Adequação da proposta, considerando: a qualificação do problema, a abordagem inter e transdisciplinar do tema enfocado, o grau de originalidade científica e tecnológica, a exequibilidade e a relevância para o desenvolvimento nacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 8.0

Critério:
Alinhamento com os temas estratégicos prioritários indicados nesta Chamada.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 2.0 **Nota:** 9.0

Critério:
Capacidade instalada das instituições integrantes do INCT para atuação em rede de pesquisa frente aos objetivos pretendidos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 8.0

Critério:
Caráter competitivo e mobilizador da proposta considerando as parcerias institucionais estabelecidas (apoio de agências de fomento, empresas e instituições do terceiro setor etc.).

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 7.0

Critério:
Capacidade e experiência do coordenador e da equipe de pesquisadores participantes em relação ao atingimento dos objetivos e metas propostos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 8.0

Critério:
Abrangência e sinergia das atividades e dos atores envolvidos na proposta, consideradas a complexidade dos temas abordados, a complementaridade de suas competências e a necessária abordagem multidisciplinar para a solução de problemas complexos, incluindo o setor empresarial e sociedade.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 7.0

Critério:
Adequação do orçamento e coerência do cronograma de execução em relação às metas e objetivos estabelecidos na proposta.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 6.0

Critério:
Estrutura operacional e modelo de gestão do INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 0.5 **Nota:** 8.5

Critério:
Adequação do Plano para Disseminação do conhecimento gerado pelo INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 0.5 **Nota:** 8.0

Critério:
Grau de inovação e potencial de impacto dos resultados esperados sob o ponto de vista científico, tecnológico, econômico e socioambiental no contexto nacional e internacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Peso: 1.0 **Nota:** 7.5
Nota Final

Nota	Ordem	Prioridade

Resultado da Avaliação
Recomendada
Justificativa:
A proposta do projeto tem uma abordagem multidisciplinar ao integrar sociolinguística, psicolinguística e tecnologia de processamento de dados linguísticos para treinar modelos de inteligência artificial. O objetivo é representar a diversidade linguística do Brasil e desenvolver uma plataforma de IA brasileira. Contudo, a proposta enfrenta várias críticas, principalmente pela falta de contextualização no cenário atual da pesquisa em IA e linguística no Brasil. Essa desconexão pode comprometer a eficácia do projeto em atender às necessidades reais do país. Um dos principais pontos negativos é a indefinição clara do escopo linguístico, que deveria especificar os tipos de dados dos componentes fonológicos, morfológicos, sintáticos, semânticos a serem coletados, além de como esses dados serão transcritos, anotados, etiquetados e sobretudo como serão disponibilizados. A proposta também não aborda adequadamente o escopo geolinguístico e as dimensões sociais e culturais das comunidades envolvidas, elementos cruciais para garantir a autenticidade e representatividade dos dados coletados. A falta de profundidade no protocolo tecnológico para implementação de algoritmos e treinamento de IA torna difícil mensurar a viabilidade do projeto em um cenário geográfico e sociolinguístico complexo, que exige mais tempo e recursos do que o estimado.Outro ponto crucial é a falta de detalhamento sobre a colaboração interdisciplinar, que pode resultar em uma abordagem fragmentada e comprometer o objetivo principal do projeto. Quanto ao alinhamento com os temas estratégicos prioritários, a proposta não identifica um problema novo em termos de gestão da diversidade linguística nacional. Embora o modelo de IA generativa proposto abranja essa diversidade, ele não resolve as disparidades regionais e socioculturais necessárias para políticas linguísticas eficazes. A metodologia ampla e genérica impede a medição concreta dos resultados esperados, e a ausência de protocolos metodológicos específicos leva a uma sensação de superficialidade.Quanto à capacidade instalada das instituições envolvidas, há uma disparidade nas áreas de especialização, com uma predominância de especialistas em sociolinguística em detrimento de áreas como psicolinguística e tecnologia. Isso limita a inovação tecnológica proposta e a capacidade de enfrentar desafios com modelos de IA. Muitos pesquisadores ainda não possuem visibilidade nacional consolidada e têm impacto limitado em políticas públicas e experiência internacional.O caráter competitivo e mobilizador da proposta também é criticado. Embora mencione parcerias internacionais, o projeto carece de impacto significativo e não demonstra uma dinâmica convincente com empresas ou

instituições do terceiro setor, o que poderia fortalecer sua implementação. A capacidade e experiência da coordenadora são adequadas em sociolinguística, mas faltam habilidades em gestão de grandes bases de dados e protocolos algorítmicos, essenciais para integração eficaz com especialistas em IA.Finalmente a abrangência e sinergia das atividades propostas são desafiadas pela complexidade do projeto, que pode ser muito ampla para o tempo previsto de execução. A coleta e organização de dados e os protocolos de treinamento de IA requerem uma abordagem mais detalhada e coordenada. O orçamento e cronograma de execução são vagos, sem detalhamento suficiente para avaliar a relevância de cada item, especialmente sem informações sobre línguas, variedades a serem investigadas, locais de coleta, entre outros aspectos práticos. Considerando a perspectiva estratégica do edital, a proposta talvez seja demasiadamente limitada à ideia de desenvolvimento de um projeto. É positivo o envolvimento de instituições menos consolidadas no país.

Recursos

Capital	Custeio	Bolsa	Valor Total
R$ 2.550.000,00	R$ 2.100.000,00	R$ 5.851.313,29	R$ 10.501.313,29

Data de Emissão

Data de Emissão do Parecer: 26/02/2025

Parecer de Ad Hoc

Critério:

Adequação da proposta, considerando: a qualificação do problema, a abordagem inter e transdisciplinar do tema enfocado, o grau de originalidade científica e tecnológica, a exequibilidade e a relevância para o desenvolvimento nacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:

Alinhamento com os temas estratégicos prioritários indicados nesta Chamada.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:

Capacidade instalada das instituições integrantes do INCT para atuação em rede de pesquisa frente aos objetivos pretendidos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Caráter competitivo e mobilizador da proposta considerando as parcerias institucionais estabelecidas (apoio de agências de fomento, empresas e instituições do terceiro setor etc.).

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Capacidade e experiência do coordenador e da equipe de pesquisadores participantes em relação ao atingimento dos objetivos e metas propostos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Abrangência e sinergia das atividades e dos atores envolvidos na proposta, consideradas a complexidade dos temas abordados, a complementaridade de suas competências e a necessária abordagem multidisciplinar para a solução de problemas complexos, incluindo o setor empresarial e sociedade.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Adequação do orçamento e coerência do cronograma de execução em relação às metas e objetivos estabelecidos na proposta.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Estrutura operacional e modelo de gestão do INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Adequação do Plano para Disseminação do conhecimento gerado pelo INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Grau de inovação e potencial de impacto dos resultados esperados sob o ponto de vista científico, tecnológico, econômico e socioambiental no contexto nacional e internacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente
Nota Final

Nota	Ordem	Prioridade

Resultado da Avaliação
Excelente
Justificativa:
Destaco, principalmente,o grau de inovação e potencial de impacto dos resultados sob o ponto de vista cinetífico, tecnológico, econômico e socioambiental no contexto nacional e internacional. Nas palavras da proponente, "A proposta do INCT promove justiça social ao garantir que comunidades indígenas, quilombolas, imigrantes e pessoas com deficiência sejam incluídas no cenário digital. Internacionalmente, o Brasil pode se posicionar como líder na criação de modelos linguísticos multiculturais e multilinguísticos, servindo de exemplo para outros países com rica diversidade cultural e linguística. Essa iniciativa não apenas enriquece a pesquisa em inteligência artificial, mas também cria uma base sólida para a redução de assimetrias globais, contribuindo para uma IA verdadeiramente inclusiva e ética."
Data de Emissão
Data de Emissão do Parecer: 02/02/2025

Parecer de Pré-seleção

Critério: O proponente é coordenador ou vice-coordendor de INCT no âmbito da Chamada INCT-CNPq nº 58/2022?
NÃO

Critério: O proponente tem seu currículo cadastrado na Plataforma Lattes até a data limite de submissão da proposta?
SIM

Critério: O proponente possui título de doutor?
SIM

Critério: O proponente é beneficiário de bolsa de Produtividade em Pesquisa (PQ), de bolsa de Produtividade em Desenvolvimento Tecnológico e Extensão Inovadora (DT) ou de bolsa PQ Sênior?
SIM

Critério: O proponente possui vinculo formal com a instituição de execução do projeto?
SIM

Critério: A instituição de execução é constituída sob as leis brasileiras e tem sua sede e administração no País?
SIM

Critério: O proponente declara no formulário de submissão que não possui qualquer inadimplência com o CNPq e com as Administrações Públicas Federal, diretas ou indiretas?
SIM

Critério: O Instituto é composto por no mínimo 8 pesquisadores doutores vinculados a, no mínimo, três instituições distintas?
SIM

Critério: Os itens financeiros solicitados na proposta são compatíveis com os permitidos na Chamada?
SIM

Nota Final

Nota	Ordem	Prioridade

Resultado da Avaliação
Enquadrada
Justificativa:
A proposta está em conformidade com o atendimento aos itens anteriores desse formulário de pré-seleção.
Data de Emissão
Data de Emissão do Parecer: 24/01/2025

Parecer de Ad Hoc

Critério:
Adequação da proposta, considerando: a qualificação do problema, a abordagem inter e transdisciplinar do tema enfocado, o grau de originalidade científica e tecnológica, a exequibilidade e a relevância para o desenvolvimento nacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Fraco

Critério:
Alinhamento com os temas estratégicos prioritários indicados nesta Chamada.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Fraco

Critério:
Capacidade instalada das instituições integrantes do INCT para atuação em rede de pesquisa frente aos objetivos pretendidos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Bom

Critério:
Caráter competitivo e mobilizador da proposta considerando as parcerias institucionais estabelecidas (apoio de agências de fomento, empresas e instituições do terceiro setor etc.).

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Fraco

Critério:
Capacidade e experiência do coordenador e da equipe de pesquisadores participantes em relação ao atingimento dos objetivos e metas propostos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Bom

Critério:
Abrangência e sinergia das atividades e dos atores envolvidos na proposta, consideradas a complexidade dos temas abordados, a complementaridade de suas competências e a necessária abordagem multidisciplinar para a solução de problemas complexos, incluindo o setor empresarial e sociedade.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Bom

Critério:
Adequação do orçamento e coerência do cronograma de execução em relação às metas e objetivos estabelecidos na proposta.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Fraco

Critério:
Estrutura operacional e modelo de gestão do INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Bom

Critério:
Adequação do Plano para Disseminação do conhecimento gerado pelo INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Fraco

Critério:
Grau de inovação e potencial de impacto dos resultados esperados sob o ponto de vista científico, tecnológico, econômico e socioambiental no contexto nacional e internacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Fraco
Nota Final

Nota **Ordem** **Prioridade**

Resultado da Avaliação
Fraco
Justificativa:
A proposta do projeto tem uma abordagem multidisciplinar ao integrar sociolinguística, psicolinguística e tecnologia de processamento de dados linguísticos para treinar modelos de inteligência artificial. O objetivo é representar a diversidade linguística do Brasil e desenvolver uma plataforma de IA brasileira. Contudo, a proposta enfrenta várias críticas, principalmente pela falta de contextualização no cenário atual da pesquisa em IA e linguística no Brasil. Essa desconexão pode comprometer a eficácia do projeto em atender às necessidades reais do país Um dos principais pontos negativos é a indefinição clara do escopo linguístico, que deveria especificar os tipos de dados dos componentes fonológicos, morfológicos, sintáticos, semânticos a serem coletados, além de como esses dados serão transcritos, anotados, etiquetados e sobretudo como serão disponibilizados. A proposta também não aborda adequadamente o escopo geolinguístico e as dimensões sociais e culturais das comunidades envolvidas, elementos cruciais para garantir a autenticidade e representatividade dos dados coletados. A falta de profundidade no protocolo tecnológico para implementação de algoritmos e treinamento de IA torna difícil mensurar a viabilidade do projeto em um cenário geográfico e sociolinguístico complexo, que exige mais tempo e recursos do que o estimado. Outro ponto crucial é a falta de detalhamento sobre a colaboração interdisciplinar, que pode resultar em uma abordagem fragmentada e comprometer o objetivo principal do projeto. O tempo previsto para a execução das atividades está possivelmente subestimado, especialmente considerando o esforço para incorporar toda a diversidade linguística brasileira. Isso cria lacunas significativas nas variedades a serem investigadas e nos critérios de seleção, não garantindo assim uma documentação representativa e de qualidade. Em relação ao alinhamento com os temas estratégicos prioritários, a proposta não consegue identificar um problema novo em termos de gestão da diversidade linguística nacional. Embora o modelo de IA generativa proposto abranja essa diversidade, ele não resolve as disparidades regionais e socioculturais necessárias para políticas linguísticas eficazes. A metodologia ampla e genérica impede a medição concreta dos resultados esperados, e a ausência de protocolos metodológicos específicos leva a uma sensação de superficialidade. Quanto à capacidade instalada das instituições envolvidas, há uma disparidade nas áreas de especialização, com uma predominância de especialistas em sociolinguística em detrimento de áreas como psicolinguística e tecnologia. Isso limita a inovação tecnológica proposta e a capacidade de enfrentar desafios com modelos de IA. Muitos pesquisadores ainda não possuem visibilidade nacional consolidada e têm impacto limitado em políticas públicas e experiência internacional. O caráter competitivo e mobilizador da proposta também é criticado. Embora mencione parcerias internacionais, o projeto carece de impacto significativo e não demonstra uma dinâmica convincente com empresas ou instituições do terceiro setor, o que poderia fortalecer sua implementação. A capacidade e experiência da coordenadora são adequadas em sociolinguística, mas faltam habilidades em gestão de grandes bases de dados e protocolos algorítmicos, essenciais para integração eficaz com especialistas em IA. Finalmente a abrangência e sinergia das atividades propostas são desafiadas pela complexidade do projeto, que pode ser muito ampla para o tempo previsto de execução. A coleta e organização de dados e os protocolos de treinamento de IA requerem uma abordagem mais detalhada e coordenada. O orçamento e cronograma de execução são vagos, sem detalhamento suficiente para avaliar a relevância de cada item, especialmente sem informações sobre línguas, variedades a serem investigadas, locais de coleta, entre outros aspectos práticos.
Data de Emissão
Data de Emissão do Parecer: 04/01/2025

Parecer de Ad Hoc

Critério:
Adequação da proposta, considerando: a qualificação do problema, a abordagem inter e transdisciplinar do tema enfocado, o grau de originalidade científica e tecnológica, a exequibilidade e a relevância para o desenvolvimento nacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Alinhamento com os temas estratégicos prioritários indicados nesta Chamada.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Capacidade instalada das instituições integrantes do INCT para atuação em rede de pesquisa frente aos objetivos pretendidos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Caráter competitivo e mobilizador da proposta considerando as parcerias institucionais estabelecidas (apoio de agências de fomento, empresas e instituições do terceiro setor etc.).

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Capacidade e experiência do coordenador e da equipe de pesquisadores participantes em relação ao atingimento dos objetivos e metas propostos.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Abrangência e sinergia das atividades e dos atores envolvidos na proposta, consideradas a complexidade dos temas abordados, a complementaridade de suas competências e a necessária abordagem multidisciplinar para a solução de problemas complexos, incluindo o setor empresarial e sociedade.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Adequação do orçamento e coerência do cronograma de execução em relação às metas e objetivos estabelecidos na proposta.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Estrutura operacional e modelo de gestão do INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Adequação do Plano para Disseminação do conhecimento gerado pelo INCT.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente

Critério:
Grau de inovação e potencial de impacto dos resultados esperados sob o ponto de vista científico, tecnológico, econômico e socioambiental no contexto nacional e internacional.

Escala de avaliação	
Excelente	A proposta aborda com sucesso todos os aspectos relevantes do critério em questão.
Muito bom	A proposta aborda satisfatoriamente o critério, embora sejam possíveis algumas melhorias.
Bom	Embora a proposta aborde amplamente o critério, há algumas deficiências.
Fraco	Existem graves deficiências em relação ao critério em questão. O critério é abordado de forma superficial e insatisfatória.
Insatisfatório	A proposta não aborda o critério em questão ou não pode ser julgada devido à falta de informações essenciais para a avaliação.

Excelente
Nota Final

Nota	Ordem	Prioridade

Resultado da Avaliação
Excelente
Justificativa:
O projeto "Plataforma da Diversidade Linguística: Dados linguísticos para uma IA brasileira" tem por objetivo organizar uma base de dados de língua atual e coletada a partir da mesma metodologia para que esta seja representativa da diversidade linguística e cultural brasileira, uma das metas do Plano Brasileiro de Inteligência Artificial (2024-2028). Assim, o projeto objetiva fornecer uma base de dados altamente qualificados com dados estruturados e etiquetados de maneira que a representatividade da nossa diversidade cultural, social e linguística contribua para fortalecer a soberania em IA. A diversidade linguística mencionada na proposta inclui, além da diversidade linguística do português brasileiro, as mais de 250 outras línguas (indígenas, de imigração, sinalizadas), negligenciadas na inclusão digital por falta de dados estruturados. A exequibilidade de um projeto tão abrangente é possível devido ao consórcio de laboratórios e grupos de pesquisa das áreas da Sociolinguística e da Psicolinguística neste INCT, contanto com profissionais com ampla e sólida experiência em suas áreas de expertise, além de outros membros de outras áreas, formando uma equipe multidisciplinar de diferentes áreas da Linguística e de outras áreas de conhecimento. O trabalho a ser desenvolvido consiste na preparação de dados linguísticos para o treinamento de modelos de língua em larga escala (LLMs), considerando a diversidade linguística brasileira, através do desenvolvimento de um protocolo conjunto de coleta de dados linguísticos em campo, a ser replicado nos grupos e laboratórios envolvidos no INCT longitudinalmente, assim como procedimentos de transcrição, alinhamento e etiquetagem de dados linguísticos para a constituição de conjunto de dados que represente a diversidade linguística brasileira. Também está incluída, nos objetivos, a realização de estudos sobre processamento linguístico da diversidade para o ajuste fino dos LLMs, contribuindo para a redução de assimetrias resultantes do treino de LLMs com traduções do inglês. O texto apresentado situa todos os itens da chamada de forma clara e consistente, atendendo de forma mais do que satisfatória aos requisitos de elaboração de um projeto de acordo com os itens da chamada CNPq no. 46/2024. Destaco a qualificação e experiência acadêmico-científica da Coordenadora e dos pesquisadores da equipe. Assim, o projeto pode contribuir para a coleta e sistematização de um conjunto de dados inclusivo e representativo da diversidade linguística brasileira com consequências positivas para políticas de inclusão digital, de aprimoramento de Inteligência Artificial, além de promover a abordagem e a discussão de questões que envolvem a interpretação de dados linguísticos e sua etiquetagem com fins de disponibilização também para pesquisas que visam o entendimento do conhecimento linguístico dos falantes de línguas naturais. Considerando então todos os pontos elencados anteriormente, sou de parecer favorável à concessão da solicitação apresentada.
Data de Emissão
Data de Emissão do Parecer: 31/12/2024