IA na Preservação da Língua Guarani: Um Estudo com Modelos Leves Inspirados no LLaMA em Comunidades Indígenas

Vinícius Tessele
Faculdade Donaduzzi
Toledo-PR, Brasil
ORCID: 0000-0002-0662-9200

Angelo Gabriel
Faculdade Donaduzzi
Toledo-PR, Brasil
ORCID: 0009-0005-7939-5301

Matheus Pinheiro Pereira
Faculdade Donaduzzi
Toledo-PR, Brasil
ORCID: 0009-0006-0915-6601

Débora Bus Cristaldo
Faculdade Donaduzzi
Toledo-PR, Brasil
ORCID: 0009-0003-9333-8034

***Abstract*—The development of Artificial Intelligence is transforming the way people interact with technology. Across the globe, there is a surge of trained models readily available for use. Faced with the challenges experienced by Indigenous communities—especially the preservation of language and cultural traditions—there is an opportunity to use AI, such as the LLaMA model, specifically lightweight models like Phi-2, Phi-3-mini, and TinyLlama, with the aim of supporting the teaching and preservation of Indigenous languages and culture. This study explores the use of large language models (LLMs) in the development of a bilingual virtual assistant (Guarani–Portuguese), dedicated to the Guarani language and traditions, with the goal of supporting the education of children and young people in the Tekoha Añetete and Tekoha Itamarã villages, located in Diamante do Oeste, Paraná, Brazil.**

***Keywords*—Artificial Intelligence; Bilingual Virtual Assistant; Guarani Language; Indigenous Education.**

***Resumo*—O desenvolvimento da Inteligência Artificial está modificando a forma como as pessoas interagem com a tecnologia. Em todos os lugares, há uma enxurrada de modelos treinados e disponíveis para uso. Diante dos desafios enfrentados pelas comunidades indígenas, em especial a preservação da língua e das tradições culturais, surge a oportunidade de utilizar a IA, como o modelo LLaMA, especificamente modelos leves como Phi-2, Phi-3-mini e TinyLlama, com o objetivo de apoiar o ensino e a preservação da língua e cultura indígena. Este estudo explora o uso de modelos de linguagem de grande porte (LLMs) no desenvolvimento de um assistente virtual bilíngue (Guarani–Português), dedicado à língua e às tradições Guarani, com o objetivo de apoiar o ensino de crianças e jovens, com foco nas aldeias Tekoha Añetete e Tekoha Itamarã, localizadas em Diamante do Oeste, Paraná.**

***Palavras-chave*—Inteligência Artificial; Assistente Virtual Bilíngue; Língua Guarani; Educação Indígena.**

I. Introdução

As aldeias Tekoha Añetete e Tekoha Itamarã, situadas em Diamante do Oeste, PR, são espaços de preservação da cultura e da língua Guarani, fundamentais para a identidade das comunidades Guarani Ñandeva e Mbya [1]. A língua Guarani é aprendida desde cedo em casa, enquanto o português é introduzido na escola a partir dos 4 ou 5 anos [2]. As escolas locais são bilíngues, com o objetivo de preservar o idioma nativo e preparar os jovens para a convivência com a sociedade mais ampla [3].

Entretanto, o acesso às tecnologias educacionais baseadas em inteligência artificial (IA) é limitado nessas comunidades, o que aumenta a desigualdade em relação às escolas urbanas [4]. Modelos de linguagem como o LLaMA podem contribuir significativamente para o registro e o ensino da língua Guarani. Iniciativas como o IndT5 [5] e o uso de corpora sintéticos são alternativas promissoras para superar a escassez de dados em línguas indígenas [6].

Diante disso, o projeto propõe o desenvolvimento de um assistente virtual bilíngue (Guarani-Português), treinado com LLaMA, para apoiar o ensino da língua Guarani e integrar o saber tradicional às tecnologias emergentes, contribuindo para a valorização e preservação cultural nas aldeias [7].

II. Referencial Teórico

Modelos de linguagem Grande (LLMs, *Large Language Models*) são modelos de redes neurais treinadas com grande volume de texto, capaz de gerar e compreender a linguagem natural com precisão [8]. Atualmente, o uso de modelos LLMs requer considerável capacidade computacional, incluindo memória RAM elevada e GPUs eficientes. Modelos leves inspirados no LLaMA, como Phi-2, Phi-3-mini e TinyLLaMA, apresentam arquitetura otimizada, exigindo menos recursos computacionais e memória, o que os torna adequados para contextos com infraestrutura limitada [9]. Esses modelos permitem a adaptação para línguas com baixa representação, mantendo desempenho razoável em tarefas de geração textual e tradução automática.

Um dos maiores desafios no uso de LLMs multilíngues é a escassez de dados em línguas minoritárias, o que limita a qualidade do treinamento e gera lacunas de desempenho em relação a idiomas amplamente representados, como o inglês. Uma das estratégias mais utilizadas para superar esse problema é a adaptação via *Low-Rank Adaptation* (LoRA), que permite ajustar modelos pré-treinados com um número reduzido de parâmetros, reduzindo os custos computacionais sem comprometer significativamente o desempenho [10].

Além do LoRA, técnicas de *continual pretraining* (pré-treinamento contínuo) com corpora específicos, geração de dados sintéticos por meio de tradução automática e *fine-tuning* supervisionado orientado por instruções (*instruction tuning*) têm se mostrado eficazes na adaptação de LLMs para contextos linguísticos de baixo recurso [11], [12]. Essas abordagens possibilitam que modelos menores, como o TinyLLaMA e o Phi-2, Phi-3-mini, sejam aplicados em línguas com pouca presença digital, mantendo relevância em tarefas de classificação, tradução e resposta a perguntas.

III. Materiais e Métodos

O projeto seguiu as etapas de pesquisa, coleta de dados, desenvolvimento de arquitetura, treinamento do modelo, construção do assistente e validação [13].

Para que possamos atingir o objetivo será necessário seguir as seguintes fases:

1) Pesquisa e fundamentação;
2) Coleta e preparação de dados;
3) Desenvolvimento da arquitetura;
4) Treinamento e *fine-tuning*;
5) Desenvolvimento do assistente;
6) Testes e validação.

A coleta de dados foi realizada com professores e alunos indígenas de uma universidade pública localizada em Foz do Iguaçu. Foram reunidas palavras e frases do cotidiano, organizadas em categorias temáticas, totalizando centenas de exemplos em Guarani. Para esse fim, desenvolveu-se uma página web interativa para capturar gravações de áudio, respeitando a oralidade da língua [14].

A etapa de construção do conjunto de dados foi orientada pelo objetivo de fornecer insumos linguísticos adequados para o treinamento e validação de modelos de linguagem voltados ao Guarani Nhandewa. A coleta concentrou-se em palavras e frases de uso cotidiano, organizadas conforme diferentes categorias temáticas e funcionais:

- 442 palavras isoladas;
- 42 frases curtas (como saudações);
- 21 frases sobre o clima;
- 15 frases com emoção;
- 136 frases completas;
- 29 perguntas simples;
- 15 comandos simples;
- 11 frases culturais e espirituais;
- 8 frases sobre a natureza e a terra.

O modelo foi treinado no Google Colab usando PyTorch e Hugging Face Transformers, com dados estruturados em arquivos JSON no formato de aprendizado supervisionado.

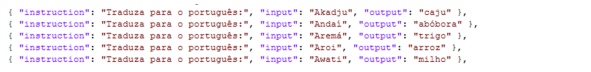

Fig. 1. Estrutura JSON

Todos os dados foram organizados em estruturas arquivos JSON Figure 1, estruturados com os campos, ”instruction”: ”Texto de instrução”, ”input”: ”texto em português”, ”output”: ”tradução em guarani”. Essa organização permitiu a utilização direta dos dados com ferramentas de pré-processamento compatíveis com modelos baseados na arquitetura Transformer. Cada par input-output representa um exemplo de aprendizado supervisionado para tarefas de geração textual. Para os experimentos, foram utilizados os seguintes modelos de linguagem de código aberto:

- Phi-2 (microsoft/phi-2): modelo com 2,7 bilhão de parâmetros, desenvolvido pela Microsoft, treinado com foco em dados de alta qualidade e conteúdos educativos.
- Phi-3-mini-4k-instruct (microsoft/Phi-3-mini-4k-instruct): Evolução do modelo Phi-2, este modelo possui 3.8 Bilhões de parâmetros e inclui técnicas de otimização de preferências humanas
- TinyLlama-1.1B-Chat-v1.0 (TinyLlama/TinyLlama-1.1B-Chat-v1.0): modelo com 1,1 bilhão de parâmetros, compatível com a arquitetura LLaMA, treinado com aproximadamente 3 trilhões de tokens.

A Figure 2 apresenta o *pipeline* completo desenvolvido para este trabalho, organizado em três etapas principais: treinamento, avaliação e geração de fala. Na fase de treinamento, o modelo recebe como insumo o conjunto de dados bilíngue em Português–Guarani, processado e normalizado para garantir consistência. Em seguida, na etapa de avaliação, o desempenho do modelo é medido por métricas como perplexidade e acurácia, permitindo ajustes finos por meio de técnicas como *fine-tuning* e LoRA.

A versão inicial do assistente oferece interação por texto e voz, tradução automática, narração de histórias e explicações culturais. Para a conversão de texto em fala (text-to-speech, TTS), foi usado o modelo facebook/mms-tts-grn, da Meta AI, permitindo síntese de voz natural em Guarani, mesmo com infraestrutura limitada. A execução do TTS foi realizada em

Entrada:
- Dataset bilíngue Guarani-Português em JSON
- Modelos LLM pré-treinados

microsoft/phi-2

microsoft/Phi-3-mini-4k-instruct

TinyLlama/TinyLlama-1.1B-Chat-v1.0
- Modelo de voz TTS para Guarani

facebook/mms-tts-grn

Etapas:
1. Carregamento e aumento de dados com instruções variadas em português e guarani
2. Formatação em estilo Instruct ("Instrução", "Entrada", "Resposta")
3. Treinamento de modelos com LoRA (PEFT) e ajuste fino leve
4. Avaliação com métricas: BLEU, ROUGE, BERTScore e Fuzzy Matching
5. Geração de respostas para prompts personalizados
6. Conversão de respostas em áudio com TTS
7. Exportação de CSVs para análise quantitativa e qualitativa

Saídas:
- Modelos LoRA treinados
- Tabelas de desempenho
- Arquivos de áudio com respostas sintetizadas

Fig. 2. Pipeline Completo: Treinamento, Avaliação e Fala

ambiente com recursos computacionais moderados, evidenciando a viabilidade do uso dessa tecnologia em contextos com limitações de infraestrutura. Não foi realizado treinamento adicional do modelo TTS, até que se consiga uma LLM que tenha resultados satisfatórios.

Por fim, a fase de fala corresponde à conversão do texto gerado em áudio, preservando a oralidade característica da língua Guarani, recurso fundamental para a valorização e transmissão da cultura indígena. Esse fluxo integrado garante tanto a robustez do modelo quanto sua aplicabilidade prática em contextos educacionais. A validação do assistente será feita por professores indígenas e universidades envolvidas, para garantir o alinhamento com os saberes e valores culturais do povo Guarani [1].

IV. Resultados

Os resultados da pesquisa demonstram a viabilidade do uso de modelos de linguagem de pequeno porte em tarefas de geração textual em Guarani, uma língua de baixa representação. Foram analisados três modelos: Phi-2, Phi-3-mini e TinyLlama. A curva de perda (loss) ao longo do treinamento revelou que o modelo TinyLlama apresentou uma queda mais rápida e uma estabilização mais eficiente, sugerindo bom desempenho em contextos com recursos computacionais limitados.

Observa-se na Figure 3 que o TinyLlama apresenta uma queda mais rápida da perda e estabilização em relação aos demais, o que sugere maior eficiência no aprendizado. Entretanto foi semelhante os resultados apresentados pelos três modelos. O modelo Phi-3-mini, embora maior em números de parâmetros,

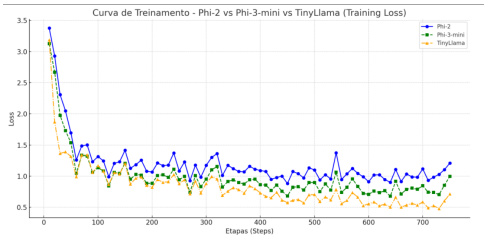

Fig. 3. Curva de Loss – Phi-2 vs Phi-3-mini vs TinyLlama

obteve também resultados consistentes. Já o modelo Phi-2, apresento maior variabilidade ao longo do treinamento. Se olharmos somente por este gráfico modelos menores, podem ser mais eficientes em tarefas de adaptação linguísticas para línguas de baixa representação, especialmente quando utilizados em ambientes com recursos computacionais limitados.

Ainda que os valores numéricos sejam baixos, o desempenho é considerado positivo, dado o desafio de lidar com uma língua pouco representada e com escassez de dados [6]. Os modelos demonstraram sinais claros de aprendizado linguístico, com destaque para o Phi-3-mini, que mostrou maior capacidade de adaptação ao contexto cultural Guarani. Os resultados detalhados podem ser observados na Tabela I:

TABELA I
DESEMPENHO DE GERAÇÃO DE TEXTO: MÉTRICAS BLEU, ROUGE, BERTSCORE E FUZZY MATCHING

Métrica	Phi-2	Phi-3-mini	TinyLlama
BLEU	0,167	0,170	0,151
ROUGE-1	0,0021	0,0212	0,0018
ROUGE-L	0,0021	0,0214	0,0018
BERTScore	0,5510	0,5714	0,5530
Fuzzy Matching	11,37%	12,21%	10,97%

Com base nos resultados obtidos Tabela 1, o modelo Phi-3-mini apresentou o melhor desempenho geral em todas as métricas. O BLEU mede a precisão lexical da saída gerada em relação a uma referência humana, neste caso o BLEU do modelo Phi-3 apresentou 0,170, indica que o modelo há aprendizado, entretanto há espaço para avanços. ROUGE – 1 compara palavras individuais e ROUGE-L considera frases e ordem de continuidade, valores baixos como 0,0212 indicam baixa sobreposição lexical entre a saída e a referência. BERTScore utiliza modelos de linguagem pré-treinados para comparar vetores de similaridades entre a saída do modelo e a referência. Um BERTScore F1 de 0.57 indica algum nível de semelhança semântica entre as respostas geradas e as esperadas. Fuzzy Matching mede a semelhança textual aproximada, um

valor de 12% indica que as respostas geradas são pouco parecidas textualmente com as esperadas.

Em relação à síntese de voz (TTS), embora o modelo facebook/mms-tts-grn tenha apresentado qualidade satisfatória na conversão de texto em áudio, sua aplicação prática foi limitada. Isso ocorreu devido ao desempenho ainda insuficiente dos modelos de linguagem na geração textual, o que comprometeu a clareza e a coerência das falas sintetizadas. Por isso, a funcionalidade de voz permanece restrita a testes experimentais até que a qualidade textual seja aprimorada.

Em síntese, os resultados apontam para o potencial de uso de modelos leves de linguagem na preservação de línguas indígenas, desde que combinados com abordagens culturalmente sensíveis e estratégias de refinamento contínuo [5].

V. Considerações Finais

Como considerações finais, é importante destacar que a pesquisa enfrentou desafios significativos, especialmente relacionados à escassez de dados e à complexidade semântica da língua Guarani, aspectos também observados por [15] em seus estudos com o Guarani-Jopara. Esses obstáculos evidenciam a dificuldade de adaptar modelos de linguagem para contextos culturais e linguísticos específicos, principalmente em línguas de baixa representação.

Apesar disso, os resultados obtidos são promissores: os modelos analisados apresentaram sinais claros de aprendizado linguístico, com destaque para o Phi-3-mini, que se mostrou o mais consistente em todas as métricas de avaliação. Isso demonstra o potencial da inteligência artificial como ferramenta de apoio à preservação cultural e linguística, desde que acompanhada de estratégias específicas de adaptação e validação junto às comunidades envolvidas [16], [17].

Referências

[1] S. C. P. Mendonça and A. F. Sella, "Atitudes linguísticas: estudo de falas de mulheres indígenas na aldeia guarani tekoha añetete," *Signum: Estudos da Linguagem*, vol. 22, no. 2, pp. 92–113, 2019, acesso em: 18 maio 2025. [Online]. Available: https://ojs.uel.br/revistas/uel/index.php/signum/article/view/38087

[2] FUNAI – Fundação Nacional do Índio, *Lições de gramática Nhandewa/Tupi-Guarani*, L. D. de Oliveira Djatsy *et al.*, Eds. Brasília, DF: FUNAI, 2018.

[3] IBGE – Instituto Brasileiro de Geografia e Estatística, "Censo demográfico 2022: população residente em territórios indígenas por sexo e idade – resultados do universo. tabela 03," Rio de Janeiro, 2023, acesso em: 18 maio 2025. [Online]. Available: https://ftp.ibge.gov.br/Censos/Censo_Demografico_2022/Quilombolas_e_Indigenas_por_sexo_e_idade_Resultados_do_universo/Tabelas_de_resultados/ods/Tabela_de_resultado_03.ods

[4] S. Isotani *et al.*, "Chatgpt pode ser aliado no processo de ensino-aprendizagem, avalia especialista," acesso em: 01 maio 2025. [Online]. Available: https://agencia.fapesp.br/chatgpt-pode-ser-aliado-no-processo-de-ensino-aprendizagem-avalia-especialista/40862

[5] E. M. B. Nagoudi, W.-R. Chen, M. Abdul-Mageed, and H. Cavusoglu, "Indt5: A text-to-text transformer for 10 indigenous languages," 2021, acesso em: 01 maio 2025. [Online]. Available: https://arxiv.org/abs/2104.07483

[6] A. Lucas, A. Baladón, V. Pardiñas, M. Agüero-Torales, S. Góngora, and L. Chiruzzo, "Grammar-based data augmentation for low-resource languages: The case of guarani-spanish neural machine translation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 1. Mexico City: Association for Computational Linguistics, 2024, pp. 6385–6397, acesso em: 01 maio 2025. [Online]. Available: https://aclanthology.org/2024.naacl-long.354.pdf

[7] L. T. Mota and V. S. de Assis, *Populações indígenas no Brasil: histórias, culturas e relações interculturais*. Maringá, Brasil: EDUEM, 2008, acesso em: 18 maio 2025. [Online]. Available: http://old.periodicos.uem.br/~eduem/novapagina/?q=node/531

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020, acesso em: 01 maio 2025. [Online]. Available: https://arxiv.org/abs/2005.14165

[9] H. Touvron, T. Lavril, G. Izacard, E. Grave, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023, acesso em: 01 maio 2025. [Online]. Available: https://arxiv.org/abs/2302.13971

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021, acesso em: 01 maio 2025. [Online]. Available: https://arxiv.org/abs/2106.09685

[11] I. Musa, T. Abebe, S. Ndlovu, and J. Adeola, "Lugha-llama: Large language models for african languages," *arXiv preprint arXiv:2504.06536*, 2025, acesso em: 01 maio 2025. [Online]. Available: https://arxiv.org/abs/2504.06536

[12] M. Khan, S. Ahmed, A. Rehman, and F. Ali, "Urdullama 1.0: A pretrained language model for the urdu language," *arXiv preprint arXiv:2502.16961*, 2025, acesso em: 01 maio 2025. [Online]. Available: https://arxiv.org/abs/2502.16961

[13] M. de Souza Oliveira and A. M. A. Maciel, "Proposta de arquitetura utilizando o paradigma soa para o avatar educação," *Revista de Engenharia e Pesquisa Aplicada*, vol. 3, no. 1, 2016.

[14] FUNAI – Fundação Nacional do Índio, *Lições de gramática Nhandewa-Guarani*, C. Marcolino *et al.*, Eds. Campinas, SP: Curt Nimuendajú, 2016.

[15] R. Díaz, D. Díaz, E. Díaz, M. Ruiz-Olazar, and M. A. Torales, "Building a large language model for guarani-jopara? methodology, challenges, and preliminary results," 2025, acesso em: 2 maio 2025. [Online]. Available: https://doi.org/10.13140/RG.2.2.32323.52006

[16] B. Deferia, "Uso y aplicación de la inteligencia artificial en el entorno educativo indígena," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 8, no. 4, pp. 10 805–10 812, 2024, acesso em: 01 maio 2025. [Online]. Available: https://doi.org/10.37811/cl_rcm.v8i4.13227

[17] A. C. B. Cabral, "Os guarani – o tempo das andanças acabou? conflitos entre ficar e partir," Dissertação (Mestrado em Ciências Sociais), Universidade Estadual do Oeste do Paraná, Toledo, 2016, acesso em: 17 maio 2025. [Online]. Available: https://tede.unioeste.br/handle/tede/2035