UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Lenon Anthony de Souza da Silva

Small Language Models for Augmentative and Alternative Communication

Recife

2025

Lenon Anthony de Souza da Silva

Small Language Models for Augmentative and Alternative Communication

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: André Câmara Alves do Nascimento

Recife

2025

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Bibliotecário(a): Ana Catarina Macêdo – CRB-4 1781

S586s Silva, Lenon Anthony de Souza da.
Small language models for augmentative and alternative communication / Lenon Anthony de Souza da Silva. – Recife, 2026.
56 f.; il.

Orientador(a): André Câmara Alves do Nascimento.

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, BR-PE, 2026.

Inclui referências e apêndice(s).

1. Dispositivos de comunicação para pessoas com deficiência. 2. Distúrbios da linguagem. 3. Inteligência artificial . 4. Sistemas de comunicação sem fio I. Nascimento, André Câmara Alves do, orient. II. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Lenon Anthony de Souza da Silva às 14:00 do dia 03/02/2026, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **Small Language Models for Augmentative and Alternative Communication**, orientado por André Câmara Alves do Nascimento e aprovado pela seguinte banca examinadora:

_____

André Câmara Alves do Nascimento
DC/UFRPE

_____

Rafael Ferreira Leite de Mello
DC/UFRPE

À minha família, que apoiaram imensamente a minha educação, batalharam desde sempre e valorizaram os estudos. Por nunca termos desistido, mesmo em momentos difíceis.

Agradecimentos

Agradeço primeiramente à minha família, por todo o apoio que me deram, mesmo quando pensei em desistir do curso no primeiro período, vocês foram a minha base e me ampararam nos momentos mais difíceis, foi assim no curso e é assim na vida. Também agradeço a Andreza, que sempre esteve do meu lado em momentos cruciais da minha vida, por sempre ter me incentivado e acreditado em mim, sou muito sortudo de ter você em minha vida. Eu os amo para todo o sempre.

Quero agradecer meus amigos dessa jornada, que fizeram a faculdade ser uma das melhores experiências que tive, como também meus amigos de ensino médio, onde também fui muito feliz.

Gostaria também de agradecer aos meus professores, que me acompanharam nessa jornada e muito me ensinaram. Obrigado a todos que me deram oportunidade de poder crescer, melhorar e exercer minha profissão a qual tenho muito carinho. Destaco meu professor e orientador André Câmara, que muito me apoiou e acreditou em mim. Também destaco Péricles, professor que me introduziu ao mundo da programação, com muito carinho, e Rafael Mello, que me introduziu ao mundo científico. Cito o professor George Valença por me fazer ter senso crítico em relação a computação e pensar em inovação, e a Taciana Pontual por me ensinar que computação e a acessibilidade conversam entre si.

De coração, agradeço por todo o esforço que todos os meus professores, familiares e amigos fizeram por mim.

Por último, e não menos importante, agradeço a Deus pela vida, por todas as oportunidades, por nunca ter deixado faltar a fé e por sempre iluminar meu caminho.

"A persistência é o caminho do êxito."
(Charles Chaplin)

Resumo

A Comunicação Aumentativa e Alternativa (CAA) é fundamental para milhões de pessoas com necessidades complexas de comunicação. Ferramentas tradicionais de CAA enfrentam um *trade-off* entre expressividade e eficiência, enquanto soluções baseadas em *Large Language Models* (LLMs) exigem conectividade e levantam preocupações de privacidade. Este trabalho investiga a especialização de *Small Language Models* (SLMs) para geração de cartões de comunicação em português brasileiro, em que cada cartão é composto por uma frase curta, uma frase longa e um símbolo visual (emoji Unicode). Foi desenvolvido um *pipeline* de construção de *dataset* combinando curadoria manual, aumento sintético via GPT-4o-mini e integração com a base ARASAAC, resultando em aproximadamente 17.800 exemplos anotados. Uma contribuição central é o *framework* de classificação baseado na distinção entre vocabulário Core (gramatical) e Fringe (tópico), fundamentado na literatura clínica de CAA. Sete modelos de três famílias de SLMs (Qwen, Llama, Gemma) foram avaliados no conjunto de teste com configuração padronizada de PEFT/LoRA e quantização 4-bit, utilizando BLEU, ROUGE e similaridade semântica sobre a string completa gerada em comparação com a referência do *dataset*. Os resultados indicam que o Qwen3-1.7B apresenta o melhor equilíbrio entre qualidade (BLEU: 0.1453, ROUGE-L F1: 0.3142, similaridade semântica: 0.77) e eficiência, viabilizando inferência local em GPUs de médio-alto desempenho com VRAM de 12GB. Foi também desenvolvida uma infraestrutura de avaliação com usuários reais, cujo piloto está planejado como próxima fase da pesquisa. Os resultados indicam a viabilidade de assistentes de CAA privados e *offline* baseados em SLMs especializados.

Palavras-chave: Comunicação Aumentativa e Alternativa. Modelos de Linguagem Compactos. *Fine-Tuning*. LoRA. Quantização. Vocabulário Core/Fringe.

Abstract

Augmentative and Alternative Communication (AAC) is essential for millions of individuals with complex communication needs. Traditional AAC tools face a trade-off between expressivity and efficiency, while Large Language Model-based solutions require connectivity and raise privacy concerns. This work investigates the specialization of Small Language Models (SLMs) for communication card generation in Brazilian Portuguese, where each card is represented by a short phrase, a long phrase, and a visual symbol (Unicode emoji). We developed a dataset construction pipeline combining manual curation, synthetic augmentation via GPT-4o-mini, and ARASAAC integration, resulting in approximately 17,800 annotated examples. A central contribution is the classification framework based on the distinction between Core (grammatical) and Fringe (topical) vocabulary, grounded in clinical AAC literature. Seven models from three SLM families (Qwen, Llama, Gemma) were evaluated on the test set with standardized PEFT/LoRA configuration and 4-bit quantization, using BLEU, ROUGE, and semantic similarity computed over the full generated string compared to the dataset reference. Results demonstrate that Qwen3-1.7B achieves the best balance between quality (BLEU: 0.1453, ROUGE-L F1: 0.3142, semantic similarity: 0.77) and efficiency, enabling local inference on mid-to-high performance GPUs with 12GB VRAM. A user evaluation infrastructure was also developed, with a pilot study planned as the next research phase. Findings indicate the viability of privacy-preserving, offline AAC assistants based on specialized SLMs.

Keywords: Augmentative and Alternative Communication. Small Language Models. Fine-Tuning. LoRA. Quantization. Core/Fringe Vocabulary.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AAC	*Augmentative and Alternative Communication* (Comunicação Aumentativa e Alternativa)
CAA	Comunicação Aumentativa e Alternativa
ARASAAC	Portal Aragonês de Comunicação Aumentativa e Alternativa
BLEU	*Bilingual Evaluation Understudy*
ELA	Esclerose Lateral Amiotrófica
FP32	Ponto Flutuante de 32 bits
GPU	*Graphics Processing Unit*
INT4	Inteiro de 4 bits
INT8	Inteiro de 8 bits
LCS	*Longest Common Subsequence*
LLM	*Large Language Model*
LoRA	*Low-Rank Adaptation*
NF4	*Normal Float 4*
NLI	*Natural Language Inference*
NLP	*Natural Language Processing*
PEFT	*Parameter-Efficient Fine-Tuning*
QLoRA	*Quantized Low-Rank Adaptation*
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
S100	Escala de 0–100 para anotação de relevância
SFT	*Supervised Fine-Tuning*
SGD	*Speech Generating Device*
SLM	*Small Language Model*
VRAM	*Video Random Access Memory*

Sumário

1 Introdução

A Comunicação Aumentativa e Alternativa (CAA) compreende um conjunto de estratégias, técnicas e ferramentas utilizadas para apoiar ou substituir a comunicação oral de pessoas com necessidades complexas de comunicação (BEUKELMAN; LIGHT, 2020). Estima-se que aproximadamente 97 milhões de pessoas no mundo apresentam condições que afetam significativamente a capacidade de comunicação verbal, incluindo transtorno do espectro autista, paralisia cerebral, afasia pós-AVC, esclerose lateral amiotrófica e atrofia muscular espinhal (BEUKELMAN; LIGHT, 2020) (dados de 2020). Para esses indivíduos, que frequentemente enfrentam comorbidades como limitações de coordenação motora, a CAA é um recurso fundamental para a interação social, familiar e autonomia comunicativa.

Historicamente, as ferramentas de CAA enfrentam um *trade-off* fundamental entre taxa de comunicação e relevância contextual (BEDROSIAN; HOAG; MCCOY, 2003; WISENBURN; HIGGINBOTHAM, 2008). Sistemas com vocabulário extenso oferecem maior capacidade expressiva e relevância contextual, mas demandam mais tempo de navegação e maior carga cognitiva, resultando em taxas de comunicação mais lentas. Por outro lado, sistemas simplificados com mensagens pré-armazenadas permitem comunicação mais rápida, porém frequentemente à custa da relevância contextual (MCCOY; BEDROSIAN; HOAG, 2001), limitando a riqueza das interações possíveis. Esse desafio de equilibrar velocidade e expressividade permanece central no *design* de sistemas de CAA (HIGGINBOTHAM et al., 2007).

Entre as abordagens mais difundidas de CAA estão os sistemas baseados em troca de figuras, como o PECS (*Picture Exchange Communication System*), desenvolvido em 1985 nos Estados Unidos, que ensina comunicação funcional por meio da troca de cartões com pictogramas – símbolos gráficos que representam conceitos, objetos ou ações (CABRAL et al., 2019). O PODD (*Pragmatic Organisation Dynamic Display*) representa uma evolução desses sistemas, organizando pictogramas de forma pragmática em pranchas estruturadas que possibilitam diversas funções comunicativas, como pedir, comentar, narrar e protestar (ESTEVES, 2023). Tais sistemas demonstram a importância central dos pictogramas como elementos visuais que permitem a comunicação não-verbal de forma intuitiva e acessível.

O surgimento de *Large Language Models* (LLMs) trouxe novas possibilidades para a CAA, com capacidade de sugerir pictogramas contextualmente relevantes e auxiliar na construção de frases (VALENCIA et al., 2023; CAI et al., 2024). No entanto, abordagens baseadas em LLMs apresentam desafios práticos: a dependência de conectividade com a internet pode limitar o uso em ambientes com acesso restrito; a latência variável de APIs externas compromete a fluidez da comunicação em tempo real; custos recorrentes de serviços em nuvem podem representar barreiras econômicas; e, criticamente, a transmissão de dados de

comunicação pessoal para servidores externos levanta preocupações legítimas sobre privacidade.

Neste contexto, os *Small Language Models* (SLMs), modelos compactos com parâmetros na ordem de centenas de milhões a poucos bilhões, emergem como alternativa. Avanços recentes em técnicas de ajuste fino eficiente, como PEFT (*Parameter-Efficient Fine-Tuning*) e LoRA (*Low-Rank Adaptation*) (HU et al., 2022; DETTMERS et al., 2023), combinados com quantização agressiva, permitem que modelos especializados executem localmente em GPUs de médio-alto desempenho (12GB VRAM), preservando a privacidade do usuário enquanto eliminam a dependência de conectividade. Repositórios de pictogramas como o ARASAAC (Portal Aragonês de Comunicação Aumentativa e Alternativa) (Gobierno de Aragón, 2024) disponibilizam um catálogo amplo com mais de 40.000 pictogramas globalmente, incluindo cerca de 13.600 entradas em português brasileiro. Deste total em português, aproximadamente 12.000 entradas foram integradas para desenvolvimento do sistema proposto.

Neste trabalho, o termo "cartão de comunicação" é utilizado como unidade de interface. A tarefa é formulada como o mapeamento de uma intenção comunicativa (texto em pt-BR) para uma saída composta por cartões, em que cada cartão contém: (i) uma frase curta, (ii) uma frase longa e (iii) um símbolo visual na forma de emoji Unicode. A avaliação considera a *string* completa gerada (saída do modelo) comparada à referência do *dataset*.

1.1 Problema de Pesquisa

Apesar do potencial dos SLMs para aplicações de CAA, lacunas significativas permanecem:

1. **Escassez de recursos em português**: Até o conhecimento dos autores, não existem *datasets* de CAA em português brasileiro com classificação vocabular adequada para treinamento de modelos de linguagem.
2. **Ausência de estudos sistemáticos**: Faltam avaliações comparativas de SLMs especializados para geração de cartões de comunicação em CAA.
3. **Requisitos de privacidade**: Dados de comunicação são sensíveis, demandando soluções que operem localmente sem transmissão de informações pessoais.

Diante dessas lacunas, este trabalho busca responder à seguinte questão de pesquisa:

> *Como especializar SLMs para sugestão de cartões de comunicação em CAA semanticamente relevantes para usuários brasileiros?*

1.2 Justificativa

A relevância deste trabalho manifesta-se em múltiplas dimensões: **Relevância Social**: A comunicação é um direito humano fundamental. Ferramentas de CAA eficientes promovem inclusão social, autonomia e qualidade de vida para indivíduos com necessidades complexas de comunicação. **Relevância Técnica**: SLMs especializados para CAA representam uma área de pesquisa emergente. Este trabalho contribui com evidências empíricas sobre a viabilidade e o desempenho dessas abordagens. **Relevância para o Português Brasileiro**: A escassez de recursos de CAA em português limita o acesso de usuários brasileiros a tecnologias assistivas de ponta. O desenvolvimento de um *dataset* anotado e de modelos especializados contribui para reduzir essa lacuna. **Privacidade por *Design***: Ao viabilizar inferência local, elimina-se a necessidade de compartilhamento de dados sensíveis de comunicação com servidores externos, atendendo a requisitos éticos e legais de proteção de dados.

1.3 Objetivos

1.3.1 Objetivo Geral

Investigar e avaliar a especialização de *Small Language Models* para geração de cartões de comunicação em Comunicação Aumentativa e Alternativa em português brasileiro, com foco em execução local, preservação de privacidade e relevância para o usuário.

1.3.2 Objetivos Específicos

1. Desenvolver um *pipeline* reprodutível para construção de *dataset* de CAA em português, integrando curadoria manual, aumento sintético e recursos do ARASAAC.
2. Propor e implementar um *framework* de classificação de vocabulário, fundamentado na literatura clínica de CAA.
3. Avaliar comparativamente os modelos especializados quanto à qualidade de geração e eficiência computacional.
4. Desenvolver infraestrutura de avaliação preparada para validação futura com usuários reais de CAA.

1.4 Organização do Trabalho

O restante deste trabalho está organizado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica, abordando conceitos de CAA, modelos de linguagem e técnicas de ajuste fino; o Capítulo 3 descreve a metodologia proposta, incluindo a construção do *dataset* e

o *framework* de classificação; o Capítulo 4 apresenta e discute os resultados experimentais; por fim, o Capítulo 5 traz as considerações finais, limitações e propostas de trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos necessários para a compreensão do trabalho. São abordados conceitos de Comunicação Aumentativa e Alternativa, modelos de linguagem e técnicas de ajuste fino eficiente.

2.1 Comunicação Aumentativa e Alternativa (CAA)

2.1.1 Conceitos Fundamentais

A Comunicação Aumentativa e Alternativa (CAA) refere-se ao conjunto de métodos, estratégias e ferramentas utilizadas para complementar ou substituir a fala de indivíduos com limitações na comunicação oral (BEUKELMAN; LIGHT, 2020). O termo "aumentativa" designa recursos que ampliam a comunicação existente, enquanto "alternativa" refere-se a sistemas que substituem completamente a fala.

Os sistemas de CAA podem ser classificados em três categorias principais (ELSAHAR et al., 2019). A **comunicação não-assistida** utiliza apenas o corpo do comunicador, incluindo gestos, expressões faciais, linguagem de sinais e vocalizações. A **comunicação assistida de baixa tecnologia** emprega recursos externos simples, como pranchas de comunicação impressas, livros de símbolos e cartões de imagens. Já a **comunicação assistida de alta tecnologia** utiliza dispositivos eletrônicos, incluindo softwares dedicados em tablets, dispositivos geradores de fala (SGDs) e aplicativos especializados como Proloquo2Go, LAMP Words for Life e TouchChat. A evolução tecnológica tem expandido significativamente as possibilidades desta última categoria, permitindo interfaces mais naturais, personalização avançada e integração com tecnologias emergentes como inteligência artificial.

2.1.2 Vocabulário Core e Fringe

Um conceito fundamental na organização de sistemas de CAA é a distinção entre vocabulário Core e Fringe, sistematicamente estudada por Baker, Hill e Devylder (2000). Esta taxonomia, consolidada na prática clínica, fundamenta a arquitetura de sistemas eficientes de comunicação.

O **vocabulário Core** (essencial) compreende palavras de alta frequência que representam aproximadamente 80% da comunicação diária, apesar de constituírem apenas 200 a 400 palavras únicas (BAKER; HILL; DEVYLDER, 2000). Embora tais proporções tenham sido estabelecidas por pesquisas em inglês, estudos linguísticos sugerem que padrões similares podem ser observados em outras línguas. Ressalta-se, contudo, que essa transferência para o

português brasileiro permanece como
textbfhipótese não validada empiricamente neste trabalho, constituindo uma limitação metodológica a ser investigada em estudos futuros com corpora de comunicação real de usuários brasileiros de CAA. Suas características fundamentais incluem composição predominantemente gramatical (verbos, pronomes, adjetivos, preposições, conjunções e palavras interrogativas), consistência entre indivíduos e ambientes de comunicação, estabilidade temporal (as palavras mais frequentes permanecem consistentes ao longo da vida) e natureza generativa, permitindo a construção flexível de sentenças diversas.

O **vocabulário Fringe** (periférico) representa os 20% restantes da comunicação, mas consiste em um número muito maior de palavras únicas. Caracteriza-se por composição predominantemente nominal, com substantivos concretos organizados por tópicos, especificidade contextual variando conforme atividades, interesses e ambientes do usuário, alta personalização incluindo nomes próprios, locais favoritos e tópicos de interesse individual, além de organização semântica que se beneficia de agrupamento por categorias temáticas. A Figura 2 e a Tabela 1 ilustram essa distinção com exemplos.

Trembath, Balandin e Togher (2007) demonstraram que, embora o vocabulário Fringe seja essencial para a especificidade comunicativa, é o vocabulário Core que possibilita a comunicação generativa. Sistemas de CAA eficientes, como o Minspeak e o LAMP Words for Life, integram ambos os tipos em layouts híbridos que combinam organização gramatical para o Core e organização semântica para o Fringe.

Figura 1 – Prancha de comunicação do sistema LAMP Words for Life® (versão bilíngue inglês/espanhol), ilustrando a organização do vocabulário Core com ícones Minspeak® em grade de 84 localizações. Fonte: PRC-Saltillo (2025).

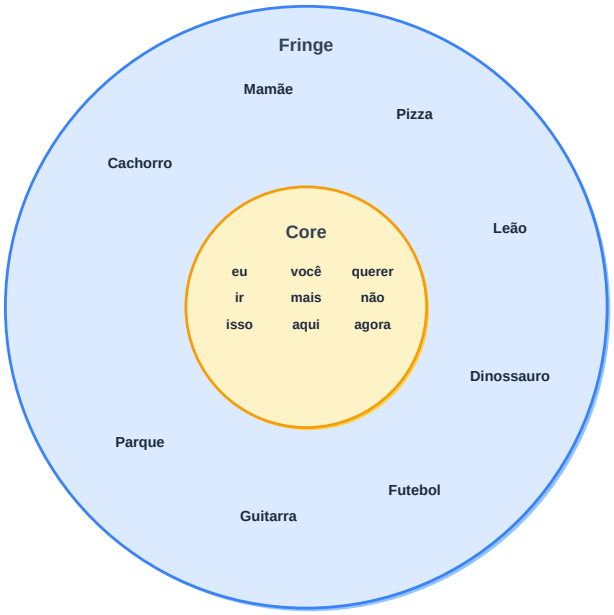

Figura 2 – Diagrama conceitual da distinção entre vocabulário Core e Fringe, com exemplos de cada tipo. Fonte: Elaboração própria.

Tabela 1 – Exemplos de vocabulário Core e Fringe

Tipo	**Categoria**	**Exemplos**
Core	Verbo	querer, fazer, ir, gostar
Core	Pronome	eu, você, ele, nós
Core	Adjetivo	bom, grande, diferente
Core	Preposição	em, para, com
Core	Interrogativo	quem, onde, quando
Fringe	Necessidades Físicas	fome, sede, dor
Fringe	Animais	cachorro, gato, pássaro
Fringe	Atividades	escola, trabalho, festa
Fringe	Objetos	telefone, livro, bola

2.1.3 Sistemas de Pictogramas: ARASAAC

O ARASAAC (Portal Aragonês de Comunicação Aumentativa e Alternativa) é um projeto do Governo de Aragão, Espanha, que disponibiliza recursos pictográficos gratuitos para CAA (Gobierno de Aragón, 2024). O sistema oferece um catálogo com mais de 40.000 pictogramas em cores e preto-e-branco, traduções para múltiplos idiomas incluindo português brasileiro (com cerca de 13.600 entradas disponíveis), licenciamento Creative Commons (CC

BY-NC-SA) permitindo uso não-comercial, organização taxonômica por categorias semânticas e classes gramaticais, além de materiais educativos e recursos para personalização. Nesta pesquisa, a integração considerou um subconjunto de aproximadamente 12.000 entradas em português brasileiro, utilizado na construção do *dataset* descrito no Capítulo 3. A estrutura do ARASAAC reflete a taxonomia multidimensional discutida anteriormente, organizando pictogramas tanto por eixo tópico (animais, alimentos, profissões) quanto por eixo gramatical (verbos, adjetivos, advérbios). Esta característica torna o ARASAAC particularmente adequado para integração com sistemas baseados em inteligência artificial.

2.2 Modelos de Linguagem

2.2.1 *Large Language Models* (LLMs)

Large Language Models são redes neurais de grande escala, tipicamente baseadas na arquitetura *Transformer* (VASWANI et al., 2017), treinadas em vastos corpora textuais para modelar distribuições de probabilidade sobre sequências de tokens. O paradigma de pré-treinamento seguido de ajuste fino possibilitou avanços significativos em diversas tarefas de processamento de linguagem natural.

A arquitetura *Transformer*, introduzida por Vaswani et al. (2017), impulsionou o campo ao substituir mecanismos recorrentes por atenção multi-cabeça (*multi-head attention*), permitindo paralelização massiva e captura eficiente de dependências de longo alcance.

Modelos como GPT-4, Claude e Gemini demonstram capacidades amplas de geração de texto, raciocínio e execução de instruções. No entanto, apresentam limitações significativas para aplicações de CAA: requisitos computacionais elevados, visto que modelos com centenas de bilhões de parâmetros exigem infraestrutura substancial; latência variável dependente de carga do servidor; dependência de conectividade estável com a internet; preocupações com privacidade, dado que dados de comunicação são transmitidos a servidores externos; e custos recorrentes associados ao uso de APIs comerciais.

2.2.2 *Small Language Models* (SLMs)

Small Language Models são modelos de linguagem compactos, com contagem de parâmetros significativamente menor que LLMs de grande escala, tipicamente na ordem de centenas de milhões a poucos bilhões (Qwen Team, 2024; DUBEY et al., 2024; Google DeepMind, 2024), projetados para equilibrar capacidade e eficiência. Avanços recentes em técnicas de ajuste fino eficiente como LoRA (HU et al., 2022) e quantização (DETTMERS et al., 2023), detalhados na Seção 2.3, têm permitido que SLMs alcancem desempenho competitivo em tarefas específicas quando adequadamente especializados.

As principais vantagens dos SLMs para aplicações de CAA incluem eficiência computacional com execução viável em GPUs de médio-alto desempenho (12GB VRAM) (Google DeepMind, 2024), preservação de privacidade através de inferência local que elimina transmissão de dados sensíveis para servidores externos, latência previsível com tempo de resposta consistente e controlável, custo reduzido pela eliminação de dependência de APIs comerciais, e capacidade de especialização através de ajuste fino eficiente para domínios específicos via técnicas PEFT (HU et al., 2022).

O *trade-off* fundamental entre tamanho de modelo e capacidade generalista pode ser mitigado através de especialização para domínios restritos. Pesquisas recentes da NVIDIA demonstram que modelos menores, adequadamente treinados para tarefas específicas, podem superar modelos maiores generalistas, particularmente em cenários de agentes autônomos e aplicações com requisitos de latência (BELCAK et al., 2025).

2.3 Técnicas de Ajuste Fino

2.3.1 PEFT e LoRA

Parameter-Efficient Fine-Tuning (PEFT) refere-se a técnicas que permitem adaptar modelos pré-treinados a tarefas específicas modificando apenas uma pequena fração dos parâmetros originais (HAN et al., 2024). Esta abordagem oferece vantagens significativas: redução drástica de requisitos de memória durante treinamento, preservação do conhecimento do modelo base, possibilidade de manter múltiplas adaptações para diferentes tarefas e viabilização de treinamento em GPUs de médio-alto desempenho.

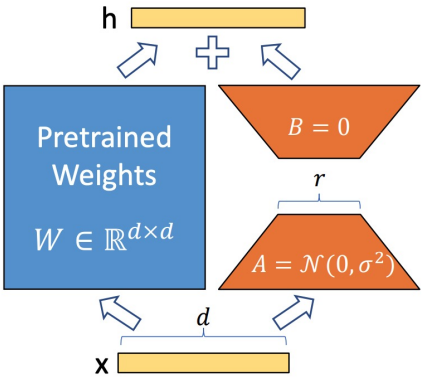

Figura 3 – Diagrama ilustrando a injeção LoRA nas camadas de atenção do *Transformer*. Fonte: Hu et al. (2022).

Low-Rank Adaptation (LoRA), proposta por Hu et al. (2022), é a técnica PEFT mais amplamente adotada. O princípio fundamental do LoRA baseia-se na hipótese de que as atualizações de peso durante o ajuste fino possuem baixa dimensionalidade intrínseca.

Formalmente, para uma camada com matriz de pesos $W_0 \in \mathbb{R}^{d \times k}$, o LoRA introduz a decomposição:

$$W = W_0 + \Delta W = W_0 + BA \tag{2.1}$$

onde $B \in \mathbb{R}^{d \times r}$ e $A \in \mathbb{R}^{r \times k}$, com $r \ll \mathsf{min}(d, k)$ sendo o *rank* da decomposição. Durante o treinamento, W_0 permanece congelado enquanto apenas A e B são atualizados.

Os hiperparâmetros principais do LoRA incluem:

- ***Rank* (r)**: Dimensionalidade da decomposição, tipicamente 8-64;
- ***Alpha* (α)**: Fator de escalonamento, geralmente $2 \times r$;
- **Módulos-alvo**: Camadas onde LoRA é aplicado (q, k, v, o, gate, up, down).

Além do LoRA, variantes e técnicas complementares têm sido propostas para diferentes cenários. O **DoRA** (*Weight-Decomposed Low-Rank Adaptation*) decompoe os pesos em magnitude e direção, aplicando LoRA apenas ao componente direcional, o que pode melhorar a capacidade de aprendizado. O **AdaLoRA** (*Adaptive Low-Rank Adaptation*) ajusta dinamicamente o rank das matrizes durante o treinamento, alocando mais parâmetros para camadas mais importantes. Neste trabalho, optou-se pelo LoRA padrão por sua maturidade, ampla adoção e compatibilidade com as bibliotecas utilizadas.

2.3.2 Quantização

Quantização refere-se à técnica de reduzir a precisão numérica dos pesos e ativações de um modelo neural. Enquanto modelos são tipicamente treinados em precisão de ponto flutuante de 32 bits (FP32), a quantização permite representação em formatos mais compactos.

Os principais formatos de quantização incluem:

- **INT8**: Inteiros de 8 bits, redução de 4× em memória;
- **INT4**: Inteiros de 4 bits, redução de 8× em memória;
- **NF4** (Normal Float 4): Formato 4-bit otimizado para distribuições normais de pesos.

Dettmers et al. (2023) propuseram o QLoRA, combinando quantização 4-bit com LoRA. Esta técnica permite o ajuste fino de modelos grandes em GPUs com memória limitada, mantendo qualidade competitiva através de quantização NF4 do modelo base, adaptadores LoRA em precisão BFloat16 e *paged optimizers* para gerenciamento de memória.

Além de técnicas voltadas à eficiência computacional (como PEFT/LoRA e quantização), é relevante distinguir **paradigmas de ajuste fino** conforme o tipo de sinal de treinamento utilizado. No *Supervised Fine-Tuning* (SFT), o modelo é especializado a partir de pares supervisionados (entrada $\rightarrow$ saída), aprendendo a reproduzir um formato alvo por meio de otimização direta da perda sobre a resposta (OUYANG et al., 2022). Em cenários em que o *dataset* já expressa a tarefa no formato desejado, o SFT tende a oferecer uma abordagem simples, reprodutível e alinhada ao objetivo de especialização.

Em contraste, técnicas de **alinhamento por preferências** utilizam comparações entre respostas como supervisão (por exemplo, "A é preferível a B"). Nesse grupo, abordagens clássicas como RLHF combinam o treinamento de um modelo de recompensa com uma etapa de otimização por aprendizado por reforço (OUYANG et al., 2022), enquanto métodos mais recentes de otimização direta por preferências (como DPO e variantes) buscam incorporar o sinal de preferência sem um laço explícito de RL, reduzindo complexidade e possíveis instabilidades (RAFAILOV et al., 2023). Há ainda cenários em que o sinal de preferência pode ser obtido de modelos auxiliares (*AI feedback*), em vez de preferências humanas, com implicações metodológicas próprias (BAI et al., 2022; LEE et al., 2023).

No escopo deste TCC, adotou-se o paradigma de SFT por compatibilidade com a estrutura do *dataset* e por simplicidade experimental, enquanto estratégias baseadas em preferências são tratadas como promissoras para etapas posteriores, especialmente por potencialmente capturarem critérios subjetivos de utilidade e adequação em contexto assistivo, desde que acompanhadas por protocolos de coleta e validação humana apropriados.

2.4 Trabalhos Relacionados

2.4.1 Modelos de Linguagem em CAA

A aplicação de modelos de linguagem em sistemas de CAA tem evoluído de abordagens baseadas em n-gramas e RNNs para arquiteturas *Transformer* de grande escala. Gupta (2024) investigaram o ajuste fino do GPT-3 (175 bilhões de parâmetros) especificamente para CAA, utilizando *domain-adaptive pretraining* e *multi-task learning*. O estudo demonstrou melhorias substanciais: redução de perplexidade de 18.2 para 12.5, economia de 42% em *keystroke savings*, e taxa de comunicação de 12 segundos por sentença. No entanto, os autores evidenciaram desafios críticos inerentes a modelos de grande escala: escassez de dados de treinamento específicos para CAA, demandas computacionais elevadas que limitam acessibilidade, e complexidade na mitigação de viés em contextos assistivos.

Valencia et al. (2023) conduziram estudo qualitativo investigando como modelos de linguagem via API podem tanto melhorar quanto prejudicar a comunicação de usuários de CAA. Os resultados revelaram que, embora LLMs ofereçam sugestões contextualmente ricas,

introduzem barreiras práticas significativas: dependência de conectividade estável, custos recorrentes de APIs comerciais, latência variável que compromete fluidez comunicativa, e preocupações sobre privacidade de dados sensíveis. O estudo destaca a tensão entre capacidade do modelo e viabilidade prática de implantação.

2.4.2 Processamento Contextual para CAA

Wisenburn e Higginbotham (2008) propuseram uma abordagem utilizando reconhecimento automático de fala do parceiro conversacional (*speaking partner*) para extrair sintagmas nominais e gerar mensagens contextualmente relevantes. O sistema Converser, combinando *speech recognition* com *parsing* de linguagem natural, alcançou ganho de 36.67% na taxa de comunicação (8.2 palavras por minuto com o sistema vs. 6.0 sem). Os autores demonstraram que o processamento do contexto conversacional em tempo real pode mitigar o *trade-off* tradicional entre taxa de comunicação e relevância contextual (BEDROSIAN; HOAG; MCCOY, 2003), permitindo produções rápidas e semanticamente apropriadas ao tópico em discussão.

2.4.3 *Transformers* para Predição em CAA

Pereira et al. (2024) desenvolveram o método PrAACT (*Predictive Augmentative and Alternative Communication with Transformers*), focado na adaptação de modelos de linguagem (como BERT e GPT-2) para a predição do próximo cartão de comunicação. Diferentemente de abordagens que exigem treinamento extensivo do zero, o método propõe a substituição da camada decodificadora do modelo por embeddings do vocabulário visual do usuário, permitindo a personalização em cenários *few-shot* e *zero-shot*.

Utilizando o corpus AACText, composto por aproximadamente 7.000 sentenças de comunicação telegráfica, o modelo adaptado (BERT-AAC) superou *baselines* pré-treinados (como o PictoBERT), alcançando uma média de AUROC (*Area Under the Receiver Operating Characteristic Curve*) de 0.807 em tarefas de ranking de relevância. A análise da distribuição de probabilidade (curtose) indicou que o método gera predições mais distribuídas e menos enviesadas que o *baseline*, sugerindo melhor generalização para vocabulários diversos.

Embora o trabalho evidencie a viabilidade de adaptar *Transformers* para CAA com baixo esforço computacional e alta flexibilidade de vocabulário, o foco da abordagem permanece na predição sequencial (autocompletar a frase cartão a cartão) dentro de um conjunto de símbolos pré-definidos. Em contraste, a presente pesquisa investiga a geração completa de cartões a partir de uma intenção comunicativa em linguagem natural.

2.4.4 Posicionamento deste Trabalho

Este trabalho diferencia-se da literatura existente em aspectos fundamentais. Quanto ao **foco em SLMs especializados para execução local**, enquanto Gupta (2024) e Valencia et al.

(2023) exploram LLMs de grande escala via API (GPT-3 com 175B parâmetros), investigamos a viabilidade de modelos compactos (poucos bilhões de parâmetros) executando localmente em GPUs de médio-alto desempenho (12GB VRAM), eliminando requisitos de conectividade e preservando privacidade. Propomos também um ***framework* Core/Fringe para NLP computacional**, sendo uma aplicação sistemática da distinção Core/Fringe da literatura clínica de CAA (BAKER; HILL; DEVYLDER, 2000) em processamento de linguagem natural, permitindo análise granular de desempenho por tipo vocabular e alinhamento com práticas fonoaudiológicas estabelecidas.

Diferentemente de Pereira et al. (2024), que focam em predição do próximo cartão em sequência, desenvolvemos capacidade de **geração contextual aberta**, produzindo múltiplos cartões de comunicação relevantes a partir de uma intenção comunicativa livre. Construímos um ***dataset* anotado para português brasileiro** com aproximadamente 17.800 exemplos e classificação Core/Fringe, expandindo recursos disponíveis para CAA em português e preenchendo lacuna identificada por trabalhos anteriores predominantemente em inglês (GUPTA, 2024; WISENBURN; HIGGINBOTHAM, 2008). A arquitetura proposta adota **privacidade por *design***, operando integralmente *offline* e atendendo requisitos éticos e legais de proteção de dados sensíveis, em contraste com soluções via API que transmitem dados do usuário para servidores externos. Por fim, realizamos **avaliação comparativa sistemática** de 7 modelos de 3 famílias de SLMs com configuração padronizada (PEFT/LoRA, quantização 4-bit), fornecendo evidências empíricas sobre *trade-offs* entre tamanho, qualidade e eficiência no contexto específico de CAA.

A Tabela 2 sintetiza as principais diferenças entre este trabalho e os trabalhos relacionados mais próximos.

Tabela 2 – Comparação entre trabalhos relacionados e este trabalho

Dimensão	**Gupta et al.**	**Pereira et al.**	**Este Trabalho**
Idioma	Inglês	Inglês	Português BR
Dataset	∼42k (públicos)	∼7k (AACText)	∼17.800 (próprio)
Modelo	GPT-3 (175B)	BERT/GPT-2	SLMs ($\leq$4B)
Técnica	*Fine-tuning* via API	*Head* substituído	LoRA + Quantização
Offline	Não	Sim	Sim
Tarefa	Predição de texto	Predição de cartão	Geração de cartões
Saída	Continuação de frase	Próximo pictograma	Cartão estruturado

3 Metodologia

Este capítulo apresenta a metodologia desenvolvida para especialização de SLMs para CAA. A Figura 4 apresenta uma visão geral do *pipeline* proposto, que compreende quatro etapas principais: (1) construção do *dataset*, (2) classificação de vocabulário Core/Fringe, (3) ajuste fino dos modelos, e (4) avaliação comparativa.

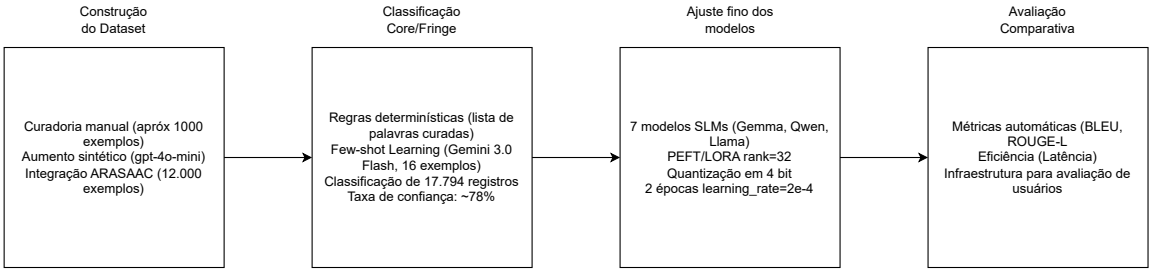

Figura 4 – Visão geral do *pipeline* proposto para especialização de SLMs para CAA.
Fonte: Elaboração própria.

3.1 Construção do *Dataset*

O *dataset* desenvolvido combina três fontes complementares, resultando em aproximadamente 17.800 exemplos únicos para treinamento.

3.1.1 Curadoria Manual

A primeira etapa consistiu na criação de um conjunto semente de aproximadamente 1.000 exemplos manualmente curados. Este conjunto foi projetado para cobrir atividades de vida diária e intenções comunicativas frequentes, incluir exemplos de vocabulário Core e Fringe balanceados, e estabelecer padrões de qualidade para aumento sintético posterior.

O formato de cada exemplo segue a estrutura:

```
Entrada: [intenção comunicativa]
Saída: [texto falado curto], [texto falado longo], [emoji]
```

Por exemplo:

```
Entrada: quero água
Saída: Água, eu quero beber água, 💧
```

3.1.2 Aumento Sintético

Para expandir o *dataset* mantendo qualidade controlada, desenvolvemos um *pipeline* de aumento sintético utilizando GPT-4o-mini através da biblioteca Distilabel (CANTO et al., 2024). O processo compreende, para cada *input* do conjunto semente, a geração de 5 sinônimos contextuais; para cada sinônimo, a geração de 5 variações de cartões de comunicação; filtragem automática por formato e qualidade; e validação amostral manual.

Os *prompts* foram cuidadosamente projetados para garantir adequação ao contexto de CAA, uso de linguagem acessível e natural, seleção de emojis semanticamente apropriados, e variação na estrutura das frases mantendo o significado.

Parâmetros de geração incluíram temperatura 1.0 (padrão da OpenAI) e *seed* 42. Ressalta-se que, em chamadas de API com temperatura não-nula, o determinismo completo não é garantido, sendo a *seed* utilizada para favorecer consistência parcial entre execuções.

3.1.3 Integração ARASAAC

A terceira fonte de dados foi o catálogo ARASAAC em português brasileiro, contribuindo com aproximadamente 12.000 entradas. O processo de integração envolveu extração de *keywords* e descrições em português, normalização de formato para o padrão do *dataset*, deduplicação exata após normalização, e validação amostral de 5% das entradas.

A Tabela 3 apresenta a composição final do *dataset*.

Tabela 3 – Composição do *dataset*

Fonte	**Quantidade**	**Percentual**
Curadoria Manual	∼1.000	5.6%
Aumento Sintético	∼4.800	27.0%
ARASAAC	∼12.000	67.4%
Total	**∼17.800**	**100%**

3.1.4 Formato do Cartão e Representação no *Dataset*

Neste trabalho, a unidade de saída é um **cartão de comunicação**, composto por três elementos: (i) uma frase curta, (ii) uma frase longa e (iii) um símbolo visual na forma de emoji Unicode. No *dataset* utilizado nos experimentos, cada cartão é representado por uma linha no formato:

```
texto_curto, texto_longo, emoji
```

Além disso, para cada intenção comunicativa, a saída contém um conjunto de $k = 5$ cartões (um por linha), refletindo múltiplas alternativas plausíveis para a mesma intenção.

3.1.5 Divisão de Dados e Protocolo Experimental

Para os experimentos de ajuste fino e avaliação, o *dataset* foi dividido em **treino/validação/teste** na proporção 70/15/15, com tamanhos: treino (12.455), validação (2.669) e teste (2.670). A avaliação automática reportada neste trabalho foi realizada exclusivamente no conjunto de teste.

3.1.6 Considerações Éticas e Privacidade

Embora o trabalho tenha como motivação o uso de CAA em contextos sensíveis, o *dataset* utilizado é composto por curadoria manual, aumento sintético e recursos do ARASAAC, não envolvendo dados pessoais de usuários reais de CAA. A infraestrutura de avaliação com usuários, quando aplicada, deve considerar consentimento informado, minimização de dados coletados e armazenamento responsável, dado o potencial caráter sensível de intenções comunicativas.

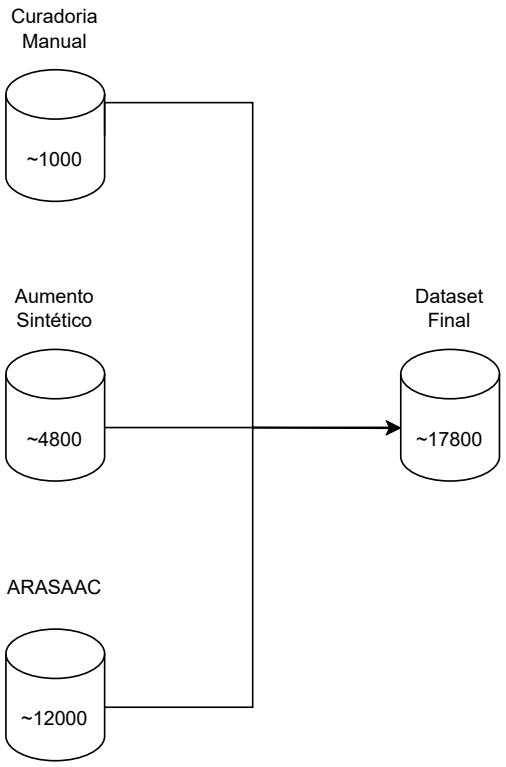

Figura 5 – Diagrama de fluxo mostrando as três fontes de dados convergindo para o *dataset* final.
Fonte: Elaboração própria.

3.2 *Framework* de Classificação Core/Fringe

3.2.1 Motivação

Em experimentos preliminares conduzidos durante o desenvolvimento deste trabalho, testou-se uma taxonomia baseada exclusivamente em categorias tópicas (animais, alimentos, vestuário, etc.), inspirada na organização semântica do próprio ARASAAC. Utilizando correspondência de palavras-chave contra listas curadas manualmente, mais de 70% dos exemplos foram atribuídos à classe residual "Outros". Isso ocorreu porque palavras funcionais de alta frequência, como verbos ("querer", "fazer"), pronomes ("eu", "você") e preposições ("para", "com"), não se encaixavam em nenhuma categoria tópica, apesar de constituírem a base da comunicação diária segundo a literatura de CAA.

A literatura de CAA há décadas reconhece que sistemas eficientes devem distinguir entre vocabulário *Core* (alta frequência, natureza gramatical) e *Fringe* (específico de contexto, natureza nominal) (BEUKELMAN; LIGHT, 2020; BAKER; HILL; DEVYLDER, 2000; TREMBATH; BALANDIN; TOGHER, 2007). No entanto, esta distinção ainda não havia sido sistematicamente aplicada em contextos de NLP computacional.

3.2.2 Taxonomia Hierárquica

Propomos uma taxonomia de dois níveis baseada na literatura clínica de CAA (BEUKELMAN; LIGHT, 2020; BAKER; HILL; DEVYLDER, 2000; TREMBATH; BALANDIN; TOGHER, 2007):

Nível 1: Tipo de Vocabulário

- **Core**: Vocabulário de alta frequência, incluindo verbos, pronomes, adjetivos, preposições, conjunções, interrogativos e interjeições.
- **Fringe**: Vocabulário periférico, composto por substantivos organizados por tópicos.

Nível 2: Subcategorias

A Tabela 4 apresenta as 16 subcategorias definidas.

Tabela 4 – Taxonomia hierárquica Core/Fringe

Tipo	**Categoria**	**Descrição e Exemplos**
Core	Verbo	Palavras de ação ou estado (querer, fazer, ir)
Core	Pronome	Referência a pessoas (eu, você, nós)
Core	Adjetivo	Palavras descritivas (grande, bom, feliz)
Core	Preposição	Palavras de posição (em, sobre, com)
Core	Conjunção	Palavras de ligação (e, mas, porque)
Core	Interrogativo	Palavras de pergunta (quem, onde, quando)
Core	Interjeição	Saudações e expressões sociais (oi, tchau, obrigado)
Fringe	Necessidades Físicas	Necessidades corporais básicas (fome, sede, dor)
Fringe	Social	Pessoas e relacionamentos (amigo, família)
Fringe	Emocional	Emoções como substantivos (amor, medo)
Fringe	Atividades	Atividades e ocupações (escola, trabalho)
Fringe	Objetos	Itens físicos (livro, telefone)
Fringe	Ambiente	Lugares e condições (casa, luz)
Fringe	Temporal	Tempo e datas (hora, dia)
Fringe	Animais	Animais (cachorro, gato)
Fringe	Outros	Conceitos não categorizados

Figura 6 – Diagrama visual da taxonomia hierárquica Core/Fringe com as 16 categorias organizadas em 2 níveis.
Fonte: Elaboração própria.

3.2.3 Metodologia de Anotação

Para classificar o *dataset* de forma escalável, desenvolvemos um *pipeline* híbrido de classificação em dois estágios, inspirado em técnicas de *weak supervision* (RATNER et al., 2017). Esta abordagem combina a precisão de regras determinísticas com a flexibilidade de LLMs para casos ambíguos.

Etapa 1: Classificação Baseada em Regras

Utilizamos regras determinísticas com listas de palavras curadas manualmente para categorias gramaticais Core bem definidas. As listas incluem palavras de alta frequência em português brasileiro para cada categoria: verbos comuns (65 verbos), pronomes (50 palavras), preposições (35 palavras), conjunções (30 palavras), interrogativos (15 palavras) e interjeições (40 expressões). Para vocabulário Fringe, aplicamos *matching* de palavras-chave em categorias semânticas como Animais, Objetos, Ambiente e Atividades. Regras com correspondência exata recebem confiança de 100%, enquanto correspondências parciais recebem 82–95%.

Etapa 2: Few-Shot Learning com LLM

Os itens não classificados no estágio anterior (confiança $< 85\%$) foram submetidos ao modelo Gemini 3.0 Flash com abordagem *few-shot learning* (BROWN et al., 2020). O *prompt* inclui 16 exemplos representativos (um para cada categoria Core e Fringe), estruturados para demonstrar o mapeamento entre entrada e categoria esperada. Estudos recentes validam o uso de LLMs para anotação em larga escala, demonstrando desempenho comparável ou superior a anotadores humanos (GILARDI; ALIZADEH; MAEGAARD, 2023).

A Tabela 5 apresenta as estatísticas do processo de classificação.

Tabela 5 – Estatísticas do *pipeline* de classificação híbrida

Métrica	**Valor**
Total de registros classificados	17.794
Classificados via regras (Etapa 1)	13.643 (76.7%)
Classificados via LLM (Etapa 2)	4.151 (23.3%)
Alta confiança ($\geq 80\%$)	14.021 (78.8%)
Confiança média geral	84.6%

3.2.4 Validação da Classificação

A aplicação do *framework* Core/Fringe com o *pipeline* híbrido resultou em cobertura de 100% dos exemplos classificados, taxa de alta confiança de 78.8%, e distribuição balanceada entre categorias Core (12.9%) e Fringe (87.1%), consistente com a literatura de CAA que indica predomínio de vocabulário específico de contexto (BANAJEE; DICARLO; STRICKLIN, 2003).

Como verificação amostral da anotação automática, foi selecionada uma amostra de 250 exemplos anotados pelo Gemini 3.0 e revisada manualmente por um único revisor humano. A comparação entre rótulos automáticos e revisão manual indicou acurácia de 99.6% para Core vs Fringe e 89.2% para a categoria (16 classes), com coeficiente Kappa de 0.992 e 0.885, respectivamente. Os valores de Kappa reportados refletem a **concordância humano-máquina**, não a confiabilidade interanotador tradicional, que exigiria múltiplos anotadores humanos independentes. Portanto, esses resultados devem ser interpretados como uma estimativa da qualidade da anotação automática, não como medida de reprodutibilidade da tarefa de anotação.

3.3 Configuração de *Fine-Tuning*

3.3.1 Modelos Base Selecionados

A seleção dos modelos base seguiu critérios de contagem de parâmetros entre 270M e 4B, capacidade multilíngue com suporte adequado ao português, compatibilidade com bibliotecas PEFT e licenciamento permissivo para pesquisa. Foram avaliados modelos de três famílias *open-source*:

Qwen (Alibaba): A família Qwen apresenta modelos com excelente desempenho multilíngue. Os modelos Qwen3 utilizados (0.6B, 1.7B e 4B) empregam arquiteturas otimizadas com suporte a raciocínio passo-a-passo (*chain-of-thought*) (Qwen Team, 2024).

Llama (Meta): A família Llama representa uma das iniciativas mais influentes em modelos abertos. O Llama 3.2-1B oferece capacidades competitivas em um formato compacto (DUBEY et al., 2024).

Gemma (Google): Desenvolvida pelo Google DeepMind, a família Gemma inclui modelos ultracompactos otimizados para eficiência. Os modelos Gemma 3 (270M) e Gemma 3n (2B e 4B) apresentam inovações arquiteturais para dispositivos com recursos limitados (Google DeepMind, 2024).

Os sete modelos selecionados são apresentados na Tabela 6.

Tabela 6 – Configuração dos modelos base

Modelo	Parâmetros	Desenvolvedor	Licença
Qwen3-0.6B	600M	Alibaba	Apache 2.0
Qwen3-1.7B	1.7B	Alibaba	Apache 2.0
Qwen3-4B-IT	4B	Alibaba	Apache 2.0
Llama-3.2-1B	1B	Meta	Llama 3
Gemma3-270M	270M	Google	Gemma
Gemma3n-e2b-it	2B	Google	Gemma
Gemma3n-e4b-it	4B	Google	Gemma

3.3.2 Configuração PEFT/LoRA

Todos os modelos foram treinados com configuração padronizada de LoRA, conforme Tabela 7.

3.3.3 Protocolo de Treinamento

O treinamento seguiu o paradigma de *Supervised Fine-Tuning* (SFT) com formato de dados de conversação de turno único (*single-turn*), *loss* calculado apenas nas respostas (*response-only training*), e utilizando Unsloth para otimização de treinamento em conjunto com TRL SFTTrainer. O hardware utilizado consistiu em GPU NVIDIA RTX 5070 (12GB

Tabela 7 – Hiperparâmetros de treinamento

Hiperparâmetro	**Valor**
LoRA *Rank* (r)	32
LoRA *Alpha* (α)	64
LoRA *Dropout*	0.0
Módulos Alvo	q, k, v, o, gate, up, down
Quantização	4-bit (NF4)
Otimizador	AdamW 8-bit
Taxa de Aprendizado	2×10^{-4}
Scheduler	Decaimento cosseno
Épocas	2
Batch Size	1–2 (por dispositivo)
Acumulação de Gradiente	4 passos
Gradient Checkpointing	Habilitado

VRAM), com *stack* de software composto por Python 3.11+, PyTorch 2.x, *Transformers* 4.47, PEFT 0.17, Unsloth e bitsandbytes para quantização 4-bit.

3.4 Métricas de Avaliação

3.4.1 Métricas de Qualidade de Geração

Para avaliação automática da qualidade de geração, utilizamos:

BLEU-4 (PAPINENI et al., 2002): Métrica de precisão de n-gramas, calculando a sobreposição entre geração e referência:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.1}$$

onde p_n é a precisão de n-gramas e BP é a penalidade de brevidade.

ROUGE-L (LIN, 2004): Baseado na subsequência comum mais longa (LCS), capturando fluência e ordenação:

$$ROUGE\text{-}L = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{3.2}$$

Similaridade Semântica: Calculada via Sentence-BERT (REIMERS; GUREVYCH, 2019), esta métrica captura equivalência de significado além da correspondência lexical. Utilizamos o modelo multilíngue `paraphrase-multilingual-MiniLM-L12-v2`, que projeta textos em um espaço vetorial de 384 dimensões. A similaridade entre a geração g e a referência r é calculada pela similaridade cosseno entre seus embeddings:

$$\text{Sim}(g, r) = \frac{\mathbf{e}_g \cdot \mathbf{e}_r}{\|\mathbf{e}_g\| \|\mathbf{e}_r\|} \tag{3.3}$$

onde $\mathbf{e}_g$ e $\mathbf{e}_r$ são os embeddings de g e r, respectivamente. Esta métrica varia de -1 a 1, sendo que valores próximos a 1 indicam alta similaridade semântica.

3.4.2 Métricas de Eficiência

Para avaliação de eficiência computacional, utilizamos latência (tempo de geração por exemplo em segundos) e uso de memória (VRAM consumida durante inferência).

3.4.3 Considerações sobre Métricas

Enfatiza-se que métricas como BLEU e ROUGE possuem limitações conhecidas para avaliação de geração de texto aberto, onde múltiplas saídas válidas existem para uma mesma entrada. Neste trabalho, BLEU/ROUGE são calculados comparando a ***string* completa** gerada pelo modelo com a referência do *dataset* (incluindo as k linhas e os três componentes de cada cartão). A similaridade semântica oferece uma perspectiva complementar, porém a validação definitiva requer avaliação com usuários reais de CAA.

4 Resultados e Discussão

Este capítulo apresenta os resultados experimentais obtidos, incluindo análise do *dataset*, desempenho comparativo dos modelos e discussão dos achados.

4.1 Análise do *Dataset*

O *dataset* final apresenta as características sumarizadas na Tabela 8. O corpus completo possui 17.794 exemplos, garantindo que cada saída siga consistentemente o formato com três componentes: frase curta, frase longa e emoji (100.0%).

Tabela 8 – Estatísticas do *dataset*

Métrica	**Valor**
Total de exemplos	17.794
Inputs únicos (não vazios)	17.792 (100.0%)
Tamanho do vocabulário	19.875 palavras
Split treino/val/teste	12.455 / 2.669 / 2.670
Formato válido (3 campos por cartão)	100.0%
Categorias	16 (7 Core + 9 Fringe)

4.1.1 Validação Manual da Anotação

Para estimar a qualidade da anotação automática realizada pelo Gemini 3.0, foi conduzida uma revisão manual amostral de 250 exemplos por um único revisor humano. Observou-se alta concordância humano-máquina para a distinção Core vs Fringe (99.6%, Kappa = 0.992) e concordância elevada para a categoria (16 classes) (89.2%, Kappa = 0.885), conforme a Tabela 9. Ressalta-se que esses valores de Kappa medem a concordância entre o LLM e o revisor humano, não a confiabilidade interanotador tradicional. Os erros residuais concentraram-se em categorias Fringe com fronteiras semânticas mais difusas, principalmente entre *Atividades* e *Outros*.

Tabela 9 – Validação manual amostral da anotação automática (n=250).

Tarefa	**Acurácia**	**Kappa**
Core vs Fringe	99.6%	0.992
Categoria (16 classes)	89.2%	0.885

4.1.2 Distribuição de Categorias

A aplicação do *framework* Core/Fringe resultou na distribuição apresentada na Tabela 10.

Tabela 10 – Distribuição por categoria

Tipo	Categoria	Quantidade	%
Core	Verbo	1.528	8.6%
Core	Adjetivo	498	2.8%
Core	Pronome	143	0.8%
Core	Interjeição	85	0.5%
Core	Preposição	25	0.1%
Core	Conjunção	15	0.1%
Core	Interrogativo	5	0.0%
Subtotal Core		**2.299**	**12.9%**
Fringe	Outros	5.755	32.3%
Fringe	Objetos	3.262	18.3%
Fringe	Atividades	2.067	11.6%
Fringe	Ambiente	1.399	7.9%
Fringe	Social	1.173	6.6%
Fringe	Necessidades Físicas	839	4.7%
Fringe	Emocional	512	2.9%
Fringe	Temporal	298	1.7%
Fringe	Animais	190	1.1%
Subtotal Fringe		**15.495**	**87.1%**
Total		**17.794**	**100%**

A distribuição mostra predominância de categorias Fringe (87.1% do total), consistente com a natureza do catálogo ARASAAC que contribui majoritariamente com substantivos concretos. As categorias Core (12.9%) fornecem o vocabulário gramatical essencial para comunicação generativa, representando verbos, pronomes e demais classes funcionais.

4.2 Desempenho dos Modelos

4.2.1 Qualidade de Geração

A Tabela 11 apresenta os resultados comparativos de qualidade de geração e eficiência para os sete modelos avaliados.

Tabela 11 – Resultados comparativos de desempenho dos SLMs na geração de cartões de comunicação

Modelo	BLEU	ROUGE-L F1	Sim. Sem.	Tempo (s)
Gemma-3-270M	0.0196	0.1280	0.6470	23.833
Gemma3n-e4b-it	0.1410	0.3075	**0.7788**	21.206
Gemma3n-e2b-it	0.1411	0.3083	0.7773	18.026
Qwen3-4B-IT	0.1221	0.2624	0.6974	12.274
Qwen3-1.7B	**0.1453**	**0.3142**	0.7724	11.751
Qwen3-0.6B	0.1125	0.2822	0.7373	6.591
Llama-3.2-1B-IT	0.1083	0.2741	0.7399	**2.547**

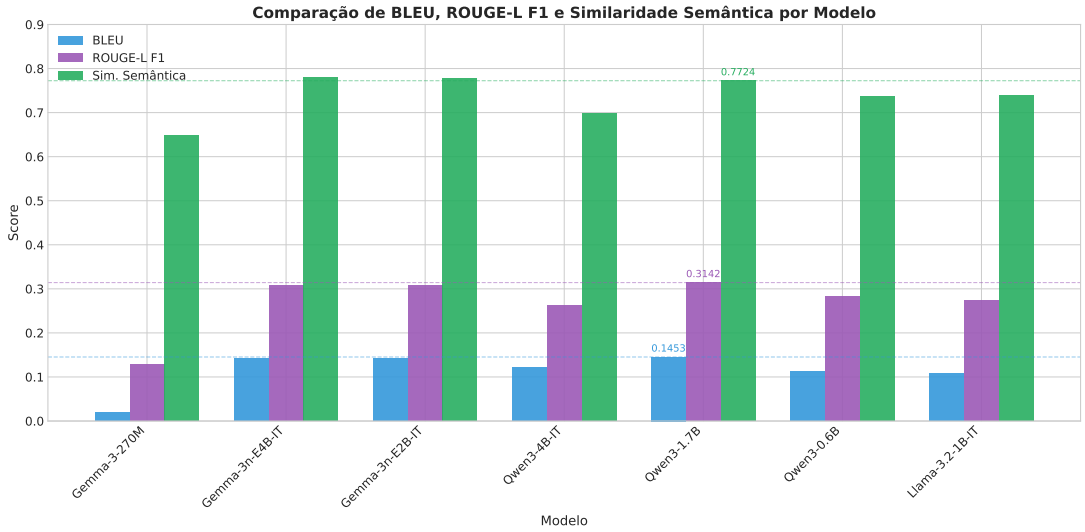

Figura 7 – Comparação de BLEU, ROUGE-L F1 e Similaridade Semântica por modelo.
Fonte: Elaboração própria.

Os resultados indicam que o **Qwen3-1.7B** alcança o melhor desempenho geral com *scores* de BLEU (0.1453) e ROUGE-L F1 (0.3142), além de similaridade semântica competitiva (0.7724). O **gemma3n-e4b-it** lidera em similaridade semântica (0.7788), seguido por gemma3n-e2b-it (0.7773). O **Llama-3.2-1B** apresenta a menor latência (2.55s), porém com qualidade inferior. No cenário avaliado, o Qwen3-4B apresenta desempenho inferior ao Qwen3-1.7B, sugerindo que o aumento de parâmetros não implica ganho proporcional após a especialização nesta tarefa. O modelo ultracompacto Gemma-3-270M apresenta desempenho inferior, com BLEU de 0.0196 e a menor similaridade semântica (0.6470).

4.2.2 Latência e Eficiência

A análise de latência revela *trade-offs* importantes. O Llama-3.2-1B oferece a melhor eficiência temporal, sendo adequado para cenários onde velocidade é crítica. O Qwen3-1.7B apresenta equilíbrio favorável entre qualidade e tempo. Já o Gemma-3-270M, apesar de ser o menor modelo, apresenta a maior latência.

4.3 Análise Comparativa

4.3.1 *Trade-off* Qualidade vs. Eficiência

A análise do *trade-off* entre qualidade e latência posiciona os modelos em três *clusters*. O primeiro, de **alta qualidade com latência moderada**, tem o Qwen3-1.7B como escolha recomendada, oferecendo o melhor equilíbrio. O segundo, de **qualidade moderada com baixa latência**, inclui Llama-3.2-1B e Qwen3-0.6B para cenários com restrições temporais estritas. O terceiro, de **qualidade moderada com alta latência**, compreende gemma3n-e2b-it e gemma3n-e4b-it, que apresentam desempenho competitivo, porém com latência elevada.

Conclusão da análise: O Qwen3-1.7B é recomendado para implantação em CAA, alcançando os melhores *scores* de qualidade com tempo de inferência razoável (∼12s por exemplo). Modelos maiores (Qwen3-4B, gemma3n-e4b-it) não demonstram ganhos proporcionais ao aumento de parâmetros.

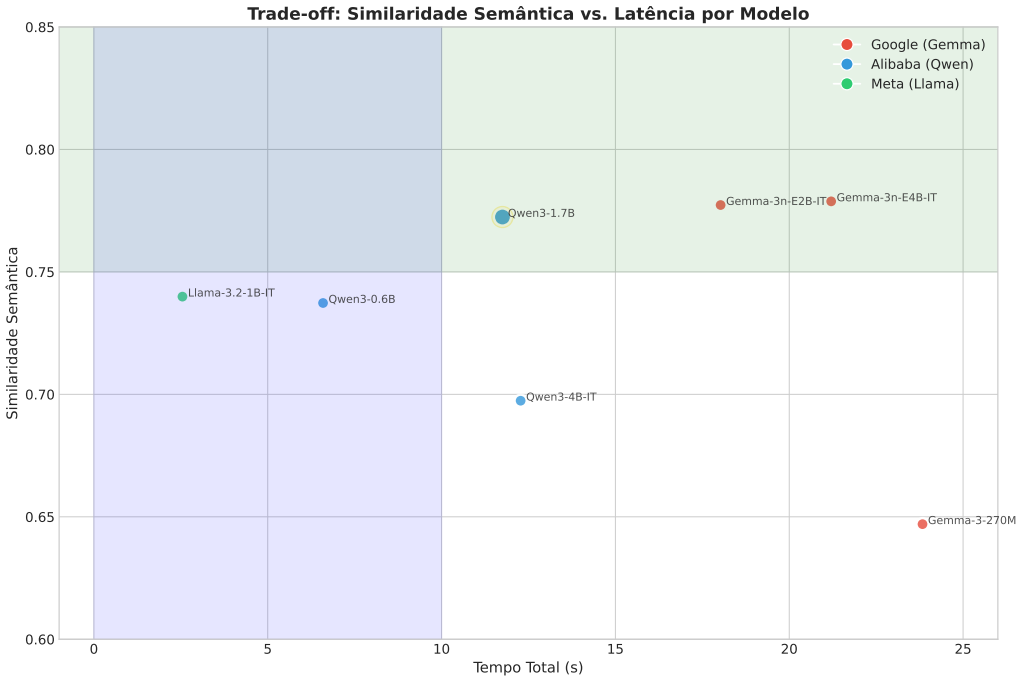

Figura 8 – *Trade-off* Similaridade Semântica vs. Latência para os 7 modelos, com destaque por fabricante. Fonte: Elaboração própria.

4.3.2 Exemplos Qualitativos de Geração

Embora as métricas lexicais como BLEU apresentem valores moderados, a similaridade semântica de aproximadamente 0.77 indica que as gerações são semanticamente próximas das referências. O BLEU penaliza variações lexicais válidas: para uma mesma intenção comunicativa, múltiplas saídas são semanticamente corretas, mas apenas uma corresponde exatamente à referência do *dataset*. Por exemplo, "eu quero beber água" e "eu quero água" são igualmente úteis para um usuário de CAA, mas a segunda receberia penalização por não coincidir lexicalmente com a primeira. Modelos bem especializados tendem a generalizar, produzindo cartões funcionalmente adequados mesmo quando lexicalmente distintos da referência.

Para ilustrar a qualidade real das gerações, a Figura 9 apresenta, para cada intenção, a referência do *dataset* (rótulos) e saídas dos dois modelos de melhor desempenho para as intenções "quero água" e "estou com dor".

Figura 9 – Exemplos qualitativos para as intenções "quero água" (acima) e "estou com dor" (abaixo): referência do *dataset* (esquerda) e cartões gerados pelos modelos Qwen3-1.7B (centro) e gemma3n-e4b-it (direita).
Fonte: Elaboração própria.

Os exemplos demonstram que ambos os modelos produzem cartões semanticamente relevantes, com frases curtas e longas adequadas ao contexto de CAA, emojis representativos e diversidade de alternativas. O gemma3n-e4b-it apresenta formulações ligeiramente mais elaboradas. Ambos os modelos geram cinco cartões por intenção, oferecendo opções variadas ao usuário.

4.4 Infraestrutura de Avaliação com Usuários

Para validação com usuários reais de CAA, foi desenvolvida uma infraestrutura completa de avaliação, inspirada metodologicamente no trabalho de Shao et al. (2018) sobre anotação para avaliação de busca de imagens.

4.4.1 Arquitetura do Sistema

O sistema compreende quatro componentes integrados: um ***frontend*** com interface web responsiva e *design* inspirado em aplicativos de CAA (Livox), apresentando cartões de comunicação em destaque; uma **API de *Feedback*** via serviço FastAPI para coleta e persistência de avaliações; um **Servidor de Inferência** com FastAPI integrando *Transformers* e Unsloth para geração em tempo real; e **PostgreSQL** para armazenamento estruturado das avaliações.

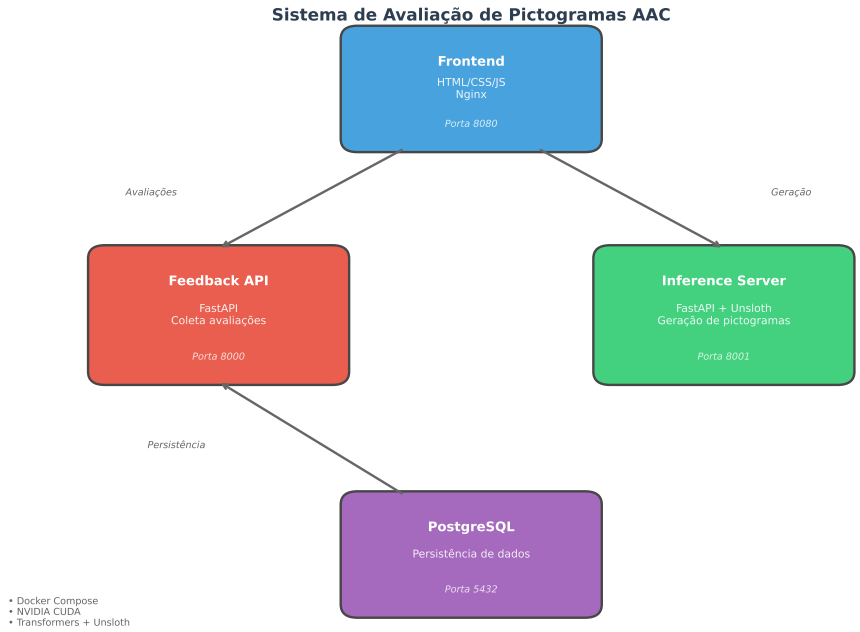

Figura 10 – Arquitetura do sistema de avaliação de cartões de comunicação em CAA, composto por *Frontend*, *Feedback API*, *Inference Server* e PostgreSQL.
Fonte: Elaboração própria.

4.4.2 Métricas de Avaliação com Usuários

Seguindo as recomendações de Shao et al. (2018), o sistema coleta múltiplas dimensões de avaliação:

1. **Relevância Tópica** (escala 0–100): Avalia o quanto os cartões gerados relacionam-se semanticamente com a intenção comunicativa do *input*. Utiliza-se escala fina (S100) conforme recomendação do estudo original.

2. **Qualidade Visual** (escala 1–4): Avalia a adequação do símbolo visual (emoji) selecionado, considerando clareza e representatividade.

3. **Coerência do Conjunto** (escala 1–3): Avalia a consistência entre os múltiplos cartões gerados para uma mesma entrada.

4. **Satisfação Geral** (escala 0–100): Métrica holística capturando a percepção global do usuário, utilizada como *gold standard* para validação das métricas componentes.

5. ***Feedback* Textual**: Campo livre para comentários qualitativos.

A escolha de escalas granulares (S100) baseia-se na evidência de Shao et al. (2018) de que anotações de granularidade fina capturam melhor as percepções dos usuários, resultando em métricas que se correlacionam mais fortemente com a satisfação.

4.4.3 Status da Avaliação

A infraestrutura encontra-se implementada e validada tecnicamente. O estudo piloto com usuários reais de CAA está planejado como próxima fase desta pesquisa.

Figura 11 – Interface do *frontend* de avaliação de cartões de comunicação, com *design* inspirado em aplicativos de CAA.
Fonte: Elaboração própria.

4.5 Discussão dos Resultados

4.5.1 Viabilidade de SLMs para CAA

Os resultados indicam que a especialização de SLMs para geração de cartões de comunicação em CAA é viável. O Qwen3-1.7B, com 1.7 bilhões de parâmetros, alcança qualidade de geração adequada operando localmente em GPUs de médio-alto desempenho (12GB VRAM).

Este achado tem implicações práticas significativas: elimina dependência de APIs comerciais, preserva privacidade de dados de comunicação, permite funcionamento *offline* e reduz custos operacionais.

4.5.2 Importância do *Framework* Core/Fringe

A taxonomia Core/Fringe desenvolvida neste trabalho constitui uma ponte entre a literatura clínica de CAA e o processamento computacional de linguagem natural. Esta contribuição é relevante por múltiplas razões.

Em primeiro lugar, o *framework* fornece uma **organização semântica fundamentada** para *datasets* de CAA. Diferentemente de categorizações puramente tópicas (que resultam em alta concentração na classe "Outros"), a distinção Core/Fringe reflete padrões reais de uso comunicativo: palavras Core (verbos, pronomes, conjunções) constituem a estrutura gramatical

da comunicação, enquanto palavras Fringe (substantivos contextuais) adicionam especificidade semântica.

Em segundo lugar, o *framework* **alinha-se com décadas de prática clínica** em CAA. Fonoaudiólogos e terapeutas organizam pranchas de comunicação seguindo esta distinção há décadas (BANAJEE; DICARLO; STRICKLIN, 2003), e a taxonomia aqui proposta operacionaliza este conhecimento clínico em formato computacionalmente tratável.

Por fim, a estrutura hierárquica de 16 categorias oferece **granularidade analítica** para futuras investigações. Pesquisadores podem avaliar desempenho segmentado por tipo de vocabulário, identificar lacunas específicas em categorias sub-representadas, e desenvolver estratégias de aumento de dados direcionadas.

4.5.3 Limitações das Métricas Automáticas

Reconhece-se que métricas como BLEU, ROUGE e similaridade semântica possuem limitações fundamentais para esta tarefa: múltiplas saídas válidas existem para uma mesma intenção comunicativa, correspondência lexical não implica adequação comunicativa, e a utilidade real para o usuário não é capturada por métricas automáticas.

Por isso, a validação definitiva requer o estudo com usuários reais, cuja infraestrutura foi desenvolvida neste trabalho.

4.5.4 Correlação Tamanho-Desempenho

Observou-se que modelos maiores não necessariamente apresentam melhor desempenho pós-especialização. O Qwen3-4B apresentou resultados inferiores ao Qwen3-1.7B, sugerindo retornos decrescentes para tamanho de modelo nesta tarefa específica, possível *overfitting* em modelos maiores com *dataset* de tamanho moderado, e importância de seleção cuidadosa do modelo base.

5 Considerações Finais

Este trabalho indicou a viabilidade de *Small Language Models* (SLMs) especializados para geração de cartões de comunicação em Comunicação Aumentativa e Alternativa (CAA) em português brasileiro. A partir do ajuste fino sistemático de sete modelos de três famílias de SLMs com técnicas PEFT/LoRA, observou-se que modelos na faixa de 1–2 bilhões de parâmetros podem alcançar qualidade competitiva operando localmente em GPUs de médio-alto desempenho, mantendo a execução *offline* como requisito central de privacidade. No conjunto de modelos avaliados, o Qwen3-1.7B destacou-se como recomendação prática por apresentar o melhor equilíbrio entre qualidade (BLEU: 0.1453, ROUGE-L F1: 0.3142, similaridade semântica: 0.77) e eficiência.

Além do desempenho, o trabalho contribui ao introduzir um *framework* de classificação e análise fundamentado na distinção entre vocabulário Core e Fringe, alinhado com a prática clínica consolidada em CAA e ainda pouco explorado na interseção com NLP computacional. Como consequência direta dessa escolha metodológica, o *dataset* produzido (aproximadamente 17.800 exemplos anotados) configura um recurso em português brasileiro com anotação vocabular sistemática, servindo de base para estudos posteriores. Por fim, a arquitetura proposta foi concebida com privacidade por *design*, eliminando a necessidade de transmitir dados sensíveis de comunicação a servidores externos e, assim, atendendo melhor a requisitos éticos e legais de proteção de dados pessoais.

5.1 Limitações do Trabalho

Reconhecem-se as seguintes limitações. Quanto ao **idioma único**, os resultados foram obtidos para português brasileiro, e a generalização para outros idiomas requer investigação específica, embora a metodologia seja transferível: as técnicas PEFT/LoRA são agnósticas ao idioma e os modelos base utilizados possuem capacidades multilíngues. Em relação à **avaliação automática**, métricas como BLEU, ROUGE e similaridade semântica, embora amplamente utilizadas, podem não capturar completamente a utilidade em contextos reais de CAA, permanecendo a correlação com satisfação de usuários como questão central. O **estudo piloto pendente** representa outra limitação: a infraestrutura de avaliação com usuários está pronta, porém a coleta de dados com usuários reais de CAA ainda não foi realizada, sendo esta validação essencial para confirmar a aplicabilidade prática.

Como verificação amostral da anotação automática, uma amostra de 250 exemplos foi revisada manualmente por um único revisor humano, indicando alta concordância humano-máquina para Core vs Fringe (99.6%, Kappa = 0.992) e concordância elevada para a categoria (89.2%, Kappa = 0.885). Ressalta-se que esses valores medem a concordância entre o LLM

anotador e o revisor humano, não a confiabilidade interanotador tradicional, que exigiria múltiplos anotadores humanos independentes. Portanto, esses resultados devem ser interpretados como uma estimativa da qualidade da anotação automática. Ademais, a categoria Fringe *Outros* concentrou 32.3% dos itens. Por restrições de tempo e recursos, não foi realizada uma revisão manual sistemática dos itens que o LLM classificou como *Outros*; inspeções pontuais indicam coexistência de conceitos que realmente extrapolam a taxonomia e itens potencialmente reclassificáveis nas categorias existentes, o que sugere necessidade de refinamento taxonômico e reclassificação supervisionada.

O **contexto conversacional limitado** também deve ser considerado, visto que a avaliação focou em geração de turno único (*single-turn*), não investigando diálogos multi-turno onde o contexto acumulado influencia a geração. Quanto ao **hardware específico**, os resultados de latência foram obtidos em GPU RTX 5070 (12GB VRAM), e a implantação em dispositivos móveis ou hardware mais restrito requer otimizações adicionais. Por fim, o **tamanho do *dataset***, embora substancial com aproximadamente 17.800 exemplos, pode limitar a capacidade de generalização, especialmente para categorias menos representadas.

5.2 Contribuições

As contribuições deste trabalho podem ser sintetizadas em cinco eixos. O primeiro é um *pipeline* de construção de *dataset*, com metodologia reprodutível, que combina curadoria manual, aumento sintético via GPT-4o-mini/Distilabel e integração com ARASAAC, com potencial de adaptação para outros idiomas e domínios. O segundo é a incorporação do *framework* Core/Fringe ao contexto de NLP, caracterizando uma aplicação sistemática da distinção Core/Fringe da literatura clínica de CAA ao processamento computacional de linguagem natural, operacionalizada por uma taxonomia de 16 categorias. O terceiro eixo é o estudo comparativo de SLMs, com avaliação padronizada de sete modelos de três famílias (Qwen, Llama e Gemma) para a tarefa específica de geração de cartões de comunicação, incluindo a análise de *trade-offs* entre qualidade e eficiência. O quarto é a entrega de uma infraestrutura de avaliação para validação com usuários reais, integrando um *frontend* com foco em acessibilidade, uma API de coleta de *feedback* e métricas inspiradas na literatura de avaliação de sistemas de informação. Por fim, o trabalho disponibiliza recursos para português brasileiro, ao reunir *dataset* anotado e modelos especializados para CAA, contribuindo para reduzir uma lacuna relevante de dados e modelos nesse idioma.

5.3 Trabalhos Futuros

Como continuidade natural, o passo mais importante é conduzir estudos com usuários reais de CAA utilizando a infraestrutura já desenvolvida, de modo a avaliar utilidade e aceitabilidade em contexto de uso e, sobretudo, investigar a correlação entre métricas auto-

máticas e satisfação para calibrar protocolos de avaliação. Em paralelo, é promissor explorar geração multimodal, integrando modelos de geração de imagem para produzir pictogramas personalizados (texto $\rightarrow$ imagem), complementando a seleção de emojis disponíveis. Outra direção relevante é a transferência *cross-lingual*, adaptando a metodologia para outros idiomas e avaliando estratégias de transferência de conhecimento entre modelos treinados em português e idiomas relacionados.

Também se destaca a personalização ao usuário, com mecanismos de adaptação contínua ao vocabulário individual, incorporando *feedback* implícito e explícito para especialização progressiva. Do ponto de vista interacional, faz-se necessário avançar para diálogos multi-turno, estendendo a avaliação a cenários em que o modelo precisa manter coerência contextual ao longo de múltiplos turnos de comunicação. Por fim, recomenda-se investigar implantação móvel, com otimizações para execução em *smartphones* e *tablets* (plataformas centrais em aplicativos de CAA), favorecendo adoção gradual em ambientes clínicos e educacionais.

Este trabalho estabelece fundamentos técnicos e metodológicos para assistentes de CAA baseados em SLMs especializados, demonstrando a viabilidade de soluções privadas, *offline* e eficientes. A continuidade natural reside na validação com usuários reais e na evolução das capacidades de personalização e multimodalidade.

Referências

BAI, Y. et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. Disponível em: <https://arxiv.org/abs/2212.08073>. Citado na página 26.

BAKER, B.; HILL, K.; DEVYLDER, R. Core vocabulary is the same across environments. In: *Proceedings of the CSUN Conference on Technology and Persons with Disabilities*. Los Angeles, CA: [s.n.], 2000. Citado 3 vezes nas páginas 19, 28 e 32.

BANAJEE, M.; DICARLO, C.; STRICKLIN, S. B. Core vocabulary determination for toddlers. *Augmentative and Alternative Communication*, Taylor & Francis, v. 19, n. 2, p. 67–73, 2003. Citado 2 vezes nas páginas 34 e 45.

BEDROSIAN, J.; HOAG, L.; MCCOY, K. Relevance and speed of message delivery trade-offs in augmentative and alternative communication. *Journal of Speech, Language, and Hearing Research*, v. 46, p. 800–817, 2003. Citado 2 vezes nas páginas 15 e 27.

BELCAK, P. et al. *Small Language Models are the Future of Agentic AI*. 2025. NVIDIA Research. Disponível em: <https://arxiv.org/abs/2506.02153>. Citado na página 24.

BEUKELMAN, D. R.; LIGHT, J. C. *Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs*. 5. ed. Baltimore, MD: Paul H. Brookes Publishing, 2020. Citado 3 vezes nas páginas 15, 19 e 32.

BROWN, T. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2020. v. 33, p. 1877–1901. Citado na página 34.

CABRAL, A. B. S. et al. O uso do PECS como tecnologia do cuidado à criança com autismo. *Revista Eletrônica Acervo Saúde*, n. 31, p. e923, 2019. Citado na página 15.

CAI, S. et al. Using large language models to accelerate communication for users with severe motor impairments. *arXiv preprint arXiv:2312.01532*, 2024. Disponível em: <https://arxiv.org/abs/2312.01532>. Citado na página 15.

CANTO Álvaro B. D. et al. *Distilabel: An AI Feedback (AIF) framework for building datasets with and for LLMs*. [S.l.]: GitHub, 2024. <https://github.com/argilla-io/distilabel>. Citado na página 30.

DETTMERS, T. et al. QLoRA: Efficient finetuning of quantized LLMs. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2023. Citado 3 vezes nas páginas 16, 23 e 25.

DUBEY, A. et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. Citado 2 vezes nas páginas 23 e 35.

ELSAHAR, Y. et al. Augmentative and alternative communication (AAC) advances: A review of configurations for individuals with a speech disability. *Sensors*, v. 19, n. 8, p. 1911, 2019. Citado na página 19.

ESTEVES, C. Do PECS ao PODD: CAA em alta tecnologia para uma criança autista. In: *Anais do X Congresso Brasileiro de Comunicação Alternativa*. [S.l.: s.n.], 2023. Citado na página 15.

GILARDI, F.; ALIZADEH, M.; MAEGAARD, M. ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023. Citado na página 34.

Gobierno de Aragón. *ARASAAC: Portal Aragonés de Comunicación Aumentativa y Alternativa*. 2024. Disponível em: <https://arasaac.org>. Acesso em: 01 nov. 2024. Citado 2 vezes nas páginas 16 e 22.

Google DeepMind. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. Citado 3 vezes nas páginas 23, 24 e 35.

GUPTA, A. *Enhancing Augmentative and Alternative Communication Systems with Fine-Tuned GPT-3: Improving Predictive Text for Users with Speech and Language Impairments*. 2024. TechRxiv Preprint. Preprint, não submetido a revisão por pares. Disponível em: <https://www.techrxiv.org/users/773740/articles/1216072>. Citado 3 vezes nas páginas 26, 27 e 28.

HAN, Z. et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. Disponível em: <https://arxiv.org/abs/2403.14608>. Citado na página 24.

HIGGINBOTHAM, D. J. et al. Access to AAC: Present, past, and future. *Augmentative and Alternative Communication*, v. 23, n. 3, p. 243–257, 2007. Citado na página 15.

HU, E. J. et al. LoRA: Low-rank adaptation of large language models. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2022. Citado 3 vezes nas páginas 16, 23 e 24.

LEE, H. et al. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*, 2023. Disponível em: <https://arxiv.org/abs/2309.00267>. Citado na página 26.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. [S.l.]: Association for Computational Linguistics, 2004. p. 74–81. Citado na página 36.

MCCOY, K.; BEDROSIAN, J.; HOAG, L. Pragmatic trade-offs in utterance-based systems: Uncovering technological implications. *ASHA Division 12 Newsletter*, p. 23–29, 2001. Citado na página 15.

OUYANG, L. et al. Training language models to follow instructions with human feedback. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2022. v. 35, p. 27730–27744. Disponível em: <https://arxiv.org/abs/2203.02155>. Citado na página 26.

PAPINENI, K. et al. BLEU: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. [S.l.]: Association for Computational Linguistics, 2002. p. 311–318. Citado na página 36.

PEREIRA, J. et al. PrAACT: Predictive augmentative and alternative communication with transformers. *Expert Systems with Applications*, v. 240, 2024. Citado 2 vezes nas páginas 27 e 28.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. Citado 2 vezes nas páginas 23 e 35.

RAFAILOV, R. et al. Direct preference optimization: Your language model is secretly a reward model. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2023. v. 36, p. 53728–53741. Disponível em: <https://arxiv.org/abs/2305.18290>. Citado na página 26.

RATNER, A. et al. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 11, n. 3, p. 269–282, 2017. Citado na página 33.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. [S.l.]: Association for Computational Linguistics, 2019. p. 3982–3992. Citado na página 36.

SHAO, S. et al. On annotation methodologies for image search evaluation. *ACM Transactions on Information Systems (TOIS)*, v. 37, n. 1, p. 1–32, 2018. Citado 2 vezes nas páginas 42 e 43.

TREMBATH, D.; BALANDIN, S.; TOGHER, L. Vocabulary selection for australian children who use augmentative and alternative communication. *Journal of Intellectual and Developmental Disability*, v. 32, n. 3, p. 191–201, 2007. Citado 2 vezes nas páginas 20 e 32.

VALENCIA, S. et al. "the less i type, the better": How AI language models can enhance or impede communication for AAC users. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. [S.l.]: Association for Computing Machinery, 2023. p. 1–14. Citado 3 vezes nas páginas 15, 26 e 28.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2017. v. 30. Citado na página 23.

WISENBURN, B.; HIGGINBOTHAM, D. J. An aac application using speaking partner speech recognition to automatically produce contextually relevant utterances: Objective results. *Augmentative and Alternative Communication*, v. 24, n. 2, p. 100–109, 2008. Citado 3 vezes nas páginas 15, 27 e 28.

Apêndices

APÊNDICE A – Prompts Utilizados

Este apêndice documenta os prompts utilizados no pipeline de construção do dataset e classificação de vocabulário. Todos os prompts foram projetados para uso com modelos de linguagem (GPT-4o-mini para geração e Gemini 3.0 Flash para classificação).

A.1 Prompt para Geração de Sinônimos

O seguinte prompt foi utilizado para gerar variações semânticas das intenções comunicativas do conjunto semente, expandindo o dataset de forma controlada.

> You are a specialized educational assistant focused on generating contextually relevant synonyms and helpful words in Brazilian Portuguese for AAC (Augmentative and Alternative Communication) systems. Your task is to analyze the input phrase or expression and generate:
>
> - 2 semantically equivalent alternatives that preserve the complete meaning and context
> - 1 random but contextually helpful Brazilian Portuguese word that could assist children with mobility challenges, AAC users, neurodivergent individuals, etc.
>
> **Critical Rule:** NEVER repeat already generated synonyms. Only generate completely new variations.
>
> For each input phrase, generate outputs that:
>
> **For the synonyms:**
>
> - Maintain the same semantic meaning as the complete input
> - Use contemporary, natural Brazilian Portuguese
> - Reflect everyday speech while maintaining clarity
> - Are appropriate for pictogram representation
> - Are easily understood by children
>
> **For the random word:**
>
> - Is related to the context or situation
> - Could be helpful for communication in similar scenarios
> - Is simple and clear
> - Is easy to represent visually

- Could expand the child's communication options

Example inputs and expected outputs:

Input: "escovar os dentes"
Output: limpar os dentes, fazer a escovação, pasta de dente

Input: "estou com fome"
Output: quero comer, preciso comer, colher

Input: "quero água"
Output: preciso beber água, estou com sede, copo

Rules for generation:

1. Use natural Brazilian Portuguese as commonly spoken today
2. Keep expressions clear and accessible while avoiding slang
3. Maintain appropriate level of formality for educational context
4. Ensure expressions are suitable for all age groups
5. Consider ease of pictogram representation
6. The random word should be useful for expanding communication options

Output format: Return only the two semantically equivalent expressions and one random helpful word as a comma-separated list in Brazilian Portuguese, without any additional text or formatting.

A.2 Prompt para Geração de Cartões de Comunicação

O seguinte prompt foi utilizado para gerar os cartões de comunicação a partir das intenções comunicativas, produzindo conjuntos de 5 cartões por entrada.

You will take the role of an expert assistant specialized in creating Augmentative and Alternative Communication (AAC) speech cards for children and individuals with various needs, including:

- Autism Spectrum Disorder (ASD)
- Speech impediments
- Motor coordination difficulties
- Developmental delays
- Communication disorders
- Non-verbal individuals

Format your response exactly as follows:

input: [word]
output: [list of options]

Rules for the output:

- Each option must have text, spoken_text, and an emoji, separated by commas
- Generate exactly 5 options following these guidelines:
  - Use simple, clear, and direct language
  - Maintain consistent sentence structures
  - Use concrete rather than abstract concepts
  - Include common daily situations
  - Ensure phrases are age-appropriate
  - Consider motor and speech limitations
  - Use positive and encouraging language
  - Avoid complex or ambiguous expressions
- Last element must be an emoji that is clearly recognizable, visually simple, directly related to the action/object, high contrast, and commonly used
- Use Brazilian Portuguese with simple grammar structures, clear pronunciation patterns, common everyday vocabulary, consistent verb tenses, and direct communication style
- Do not include counters or extra text

Focus on: Basic needs, daily routines, emotional expressions, social interactions, emergency situations, common requests, personal care, and learning activities.

Examples:

input: Ação
output: Abrir, eu quero abrir, [cadeado aberto]
Fechar, eu quero fechar, [cadeado fechado]
Ligar, eu quero ligar, [tomada]
Desligar, eu quero desligar, [tomada desligada]
Subir, eu quero subir, [seta para cima]

input: Banheiro
output: Ir ao Banheiro, eu preciso ir ao banheiro, [banheiro]
Pedir para Usar o Banheiro, eu gostaria de usar o banheiro, [vaso sanitário]
Lavar as Mãos, eu quero lavar as mãos, [sabonete]
Buscar Papel Higiênico, eu preciso de papel higiênico, [papel higiênico]
Desinfetar as Mãos, eu quero desinfetar as mãos, [frasco de loção]

A.3 Prompt para Classificação Core/Fringe

O seguinte prompt foi utilizado na etapa de classificação via LLM (Gemini 3.0 Flash) para os itens não classificados pelo sistema de regras. O prompt segue a abordagem *few-shot learning* com 16 exemplos representativos.

Classifique cada palavra na categoria Core/Fringe apropriada.

CATEGORIAS VÁLIDAS:

CORE VOCABULARY (Vocabulário Essencial - Gramatical):

1. Core: Verbo – "querer água" → Core: Verbo (ação/estado)
2. Core: Pronome – "eu" → Core: Pronome (pronome pessoal)
3. Core: Adjetivo – "grande" → Core: Adjetivo (qualidade)
4. Core: Preposição – "em cima" → Core: Preposição (posição)
5. Core: Conjunção – "porque" → Core: Conjunção (conectivo)
6. Core: Palavra de Pergunta – "onde está" → Core: Palavra de Pergunta (interrogativo)
7. Core: Interjeição – "obrigado" → Core: Interjeição (saudação/cortesia)

FRINGE VOCABULARY (Vocabulário Periférico - Semântico):

8. Fringe: Necessidades Físicas – "banheiro" → Fringe: Necessidades Físicas (necessidade corporal)
9. Fringe: Social – "amigo" → Fringe: Social (pessoa/relacionamento)
10. Fringe: Emocional – "felicidade" → Fringe: Emocional (emoção como substantivo)
11. Fringe: Atividades – "jogar futebol" → Fringe: Atividades (atividade/hobby)
12. Fringe: Objetos – "bola" → Fringe: Objetos (objeto físico)
13. Fringe: Ambiente – "escola" → Fringe: Ambiente (lugar/local)
14. Fringe: Temporal – "hoje" → Fringe: Temporal (tempo/data)
15. Fringe: Animais – "cachorro" → Fringe: Animais (animal)
16. Fringe: Outros – "acessibilidade" → Fringe: Outros (conceito abstrato)

FORMATO DE SAÍDA OBRIGATÓRIO:

1,Core: Verbo,95
2,Fringe: Objetos,88

REGRAS:

- Core = função gramatical (verbo, pronome, preposição, adjetivo, conjunção, pergunta, interjeição)

- Fringe = substantivo/conceito (necessidades, social, emocional, atividades, objetos, ambiente, temporal, animais, outros)
- Confiança = 0–100
- NÃO adicione explicações, apenas as linhas numeradas
- COMECE IMEDIATAMENTE com "1,categoria,número"