Avaliação Comparativa e Otimização de *Small Language Models* para Extração de Informações Aplicadas em Comunicação Assistiva

Caio Fontes de Melo[1], **André Câmara Alves do Nascimento**[1], **Taciana Pontual da Rocha Falcão**[1]
Valmir Macário Filho[1], **Lenon Anthony de Souza da Silva**[1]

[1] Universidade Federal Rural de Pernambuco (UFRPE)
Departamento de Computação
Recife – PE – Brasil

Abstract. *Augmentative and Alternative Communication (AAC) systems can support the autonomy of people with speech impairments by providing alternative ways to express needs, preferences, and intentions. However, many AAC applications still depend on manually predefined communication boards, which may limit flexibility and contextual adaptation. This work investigates the use of Small Language Models (SLMs) in two natural language processing tasks related to AAC: intent classification of questions and extraction of choice options from questions containing explicit alternatives. Five SLMs were fine-tuned using Quantized Low-Rank Adaptation (QLoRA) and 4-bit quantization and evaluated against a Bidirectional Encoder Representations from Transformers (BERT) baseline for intent classification and a Gemini 3.1 Pro reference for option extraction. The results show that fine-tuned SLMs can achieve competitive performance in both tasks. In particular, the BERT baseline maintained the best performance in intent classification, closely followed by Llama-3. However, Phi-3 Mini achieved the best performance in option extraction, surpassing the Gemini reference. These findings suggest that compact language models specialized for specific tasks may be a viable alternative for intelligent and efficient AAC applications.*

Keywords: Augmentative and Alternative Communication; Small Language Models; Natural Language Processing; Intent Classification; Information Extraction.

Resumo. *Sistemas de Comunicação Aumentativa e Alternativa (CAA) podem apoiar a autonomia de pessoas com deficiência de fala ao oferecerem formas alternativas de expressar necessidades, preferências e intenções. No entanto, muitas aplicações de CAA ainda dependem de pranchas de comunicação definidas manualmente, o que pode limitar a flexibilidade e a adaptação contextual. Este trabalho investiga o uso de Small Language Models (SLMs) em duas tarefas de processamento de linguagem natural relacionadas à CAA: classificação de intenção de perguntas e extração de opções de escolha em perguntas contendo alternativas explícitas. Cinco SLMs foram ajustados por meio de Quantized Low-Rank Adaptation (QLoRA) e quantização em 4 bits, sendo avaliados em comparação com um baseline baseado no Bidirectional Encoder Representations from Transformers (BERT) na tarefa de classificação de intenção e com uma referência Gemini 3.1 Pro na tarefa de extração de opções. Os resultados*

mostram que SLMs ajustados podem alcançar desempenho competitivo em ambas as tarefas. Em particular, o baseline BERT manteve o melhor desempenho na classificação de intenção, seguido de perto pelo Llama-3. No entanto, o Phi-3 Mini obteve o melhor desempenho na extração de opções, superando a referência Gemini. Esses achados sugerem que modelos de linguagem compactos e especializados para tarefas específicas podem ser uma alternativa viável para aplicações de CAA inteligentes e eficientes.

Palavras-chave: Comunicação Aumentativa e Alternativa; Modelos de Linguagem Compactos; Processamento de Linguagem Natural; Classificação de Intenção; Extração de Informação.

Submetido ao ENIAC 2026 em: 8 de junho de 2026.
TCC defendido e aprovado em: 07 de julho de 2026.

1. Introdução

A Comunicação Aumentativa e Alternativa (CAA) tem se mostrado essencial para promover a inclusão de pessoas com necessidades complexas de comunicação, oferecendo interfaces adaptadas que substituem ou complementam a comunicação verbal [1]. Em um estudo epidemiológico realizado no Reino Unido, Creer *et al.* estimaram que 536 pessoas a cada 100.000 habitantes, aproximadamente 0,5% da população, poderiam se beneficiar de recursos de CAA [2]. Esse público inclui pessoas com diferentes condições associadas a limitações comunicativas, como paralisia cerebral, autismo, acidente vascular cerebral, doença de Parkinson e demência.

Os sistemas de CAA operam fundamentalmente por meio de pranchas de comunicação que organizam símbolos gráficos. Tradicionalmente, essas ferramentas consistem em materiais físicos estáticos, como cartões plastificados, nos quais o vocabulário disponível é limitado aos símbolos previamente selecionados e só pode ser alterado manualmente. Com a evolução tecnológica, surgiram pranchas digitais dinâmicas, que permitem organizar um número maior de símbolos em telas, categorias e níveis de navegação, além de possibilitar recursos como síntese de voz e personalização da interface [1].

Embora essas aplicações modernas de CAA tenham ampliado as possibilidades de personalização e acessibilidade, muitas ferramentas ainda demandam a criação manual de pranchas de comunicação por familiares, terapeutas ou professores. Essa abordagem pode restringir as possibilidades de expressão do usuário, limitando a comunicação a conteúdos previamente definidos por terceiros. Trabalhos recentes têm investigado o uso de técnicas de IA generativa para tornar sistemas de CAA mais dinâmicos e contextuais, permitindo a geração automática de cartões e sugestões de comunicação adaptadas ao ambiente e às interações do usuário, bem como à fala transcrita dos seus interlocutores [3, 4].

Recentes avanços em *Large Language Models* (LLMs) têm ampliado as possibilidades de automação e de assistência contextual em sistemas de CAA, embora ainda existam desafios relacionados à confiabilidade e à adaptação às necessidades individuais dos usuários [5]. Além disso, a implantação de LLMs em dispositivos de borda enfrenta limitações de memória, latência e consumo de energia. Nesse contexto, técnicas de quantização e os *Small Language Models* (SLMs) têm sido investigados como alternativas para viabilizar inferência *on-device* mais leve e eficiente [6, 7].

Segundo Lu *et al.* [7], os SLMs possuem potencial para viabilizar aplicações inteligentes em dispositivos móveis, especialmente em cenários sensíveis à latência e privacidade, o que os torna particularmente relevantes para a CAA. Avanços recentes demonstram que modelos compactos, como o Phi-3, alcançam desempenho competitivo em *benchmarks* de linguagem natural, mantendo tamanho reduzido o suficiente para execução local em *smartphones* [8]. Esse cenário fortalece o potencial de aplicações assistivas baseadas em inferência local, reduzindo a dependência de servidores externos.

No contexto de aplicações assistivas e sistemas de CAA, trabalhos anteriores já investigaram o uso de Processamento de Linguagem Natural (PLN) e de Aprendizado de Máquina (AM) para reduzir barreiras comunicativas dos usuários. Emil *et al.* [9] propuseram um classificador capaz de identificar perguntas com listas de respostas, separar a pergunta das opções e apresentar pictogramas correspondentes ao usuário. A solução foi projetada para funcionar *offline* e em dispositivos com recursos limitados, alcançando redução de tempo e de interações físicas em comparação com o uso tradicional da aplicação. Contudo, por ser fundamentada em métodos heurísticos e regras manuais rígidas, essa abordagem apresenta limitações ao lidar com a variabilidade da fala natural. Para superar essas restrições, este trabalho investiga o uso de SLMs ajustados em tarefas relacionadas à CAA. A pesquisa tem como foco modelos compactos, potencialmente adequados a cenários de execução local, que são capazes de generalizar a compreenssão semântica e extrair informações estruturadas mesmo em frases não padronizadas.

Para isso, avaliamos diferentes SLMs em duas tarefas de processamento de linguagem natural relevantes para aplicações assistivas: a classificação de intenção de perguntas e a extração de opções de escolha em perguntas contendo alternativas explícitas. No contexto da CAA, identificar a intenção de uma pergunta significa reconhecer o tipo de resposta esperado pelo interlocutor (e.g., a partir da transcrição do áudio), permitindo que o sistema antecipe a interface ou os cartões mais adequadas para o usuário. Dessa forma, a tarefa de classificação focou em distinguir entre perguntas fechadas do tipo sim/não, como *"Você está com dor?"*, que podem acionar cartões rápidos de afirmação ou negação, e perguntas abertas, como *"O que você está sentindo?"*, que demandam a exibição de um vocabulário mais amplo. Na tarefa de classificação, utilizamos um modelo discriminativo baseado no *Bidirectional Encoder Representations from Transformers* (BERT) como *baseline*; na tarefa de extração de opções, incluímos o Gemini como referência de IA generativa de grande porte para contextualizar o desempenho dos modelos compactos.

Os experimentos foram conduzidos com modelos ajustados por meio de *Quantized Low-Rank Adaptation* (QLoRA) e quantização em 4 bits, sob um protocolo padronizado de treinamento e avaliação, a fim de permitir comparações consistentes entre modelos de diferentes escalas. Os resultados indicaram que SLMs ajustados podem alcançar desempenho competitivo em ambas as tarefas. Em particular, o *baseline* discriminativo manteve o melhor desempenho na classificação de intenção, seguido de perto pelo Llama-3. Por outro lado, o Phi-3 Mini apresentou desempenho superior ao da referência Gemini na extração de opções, sugerindo que modelos compactos especializados podem ser eficazes em tarefas específicas de compreensão e de geração estruturada de linguagem.

2. Materiais

Esta seção apresenta os principais materiais utilizados nos experimentos, incluindo as bases de dados utilizadas nas tarefas avaliadas e os modelos de linguagem utilizados nos processos de treinamento e inferência. Os experimentos foram conduzidos com foco na comparação de SLMs em tarefas relacionadas à CAA, segundo um protocolo de avaliação padronizado.

2.1. Bases de Dados

Para a tarefa de classificação de intenção, foi utilizada uma versão balanceada do *dataset* AmazonQA, originalmente proposto para tarefas de perguntas e respostas baseadas em avaliações de produtos. O AmazonQA reúne perguntas feitas por usuários em páginas de produtos da Amazon, associadas a respostas da comunidade e trechos de avaliações de consumidores, sendo um recurso adequado para investigar diferentes tipos de perguntas em linguagem natural [10]. A escolha dessa base foi motivada pela diversidade estrutural dessas perguntas e pela existência de metadados sobre o tipo da pergunta, o que permitiu sua adaptação para uma tarefa de classificação de intenção.

Neste trabalho, o conjunto foi adaptado para uma tarefa binária cujo objetivo é identificar o formato de resposta esperado pela pergunta, e não gerar a resposta em si. Assim, perguntas fechadas, que podem ser respondidas com sim ou não, foram associadas à classe *sim_nao*, enquanto perguntas que demandam uma resposta descritiva ou informacional foram associadas à classe *aberta*. Nesse contexto, a classe *aberta* refere-se a perguntas que exigem uma resposta informacional, mas não necessariamente apresentam alternativas explícitas no enunciado. Por exemplo, uma pergunta como *"Will these work with the Phillips sonicare handles?"* pertence à classe *sim_nao*, enquanto *"What kind of sim card it use?"* pertence à classe *aberta*. A base balanceada inicial utilizada nesta etapa contém 5.142 instâncias, sendo 2.571 perguntas da classe *sim_nao* e 2.571 perguntas da classe *aberta*. A partir dessa base, definiu-se um conjunto de teste fixo com 1.029 instâncias. Para evitar vazamento de dados, foram removidas do conjunto de treinamento as instâncias com sobreposição textual em relação ao conjunto de teste, resultando em 4.094 exemplos utilizados no treinamento dos modelos. Esse procedimento foi adotado para reduzir viés de classe e garantir uma comparação mais adequada entre o *baseline* discriminativo e os SLMs avaliados.

Para a tarefa de extração de opções, foi utilizado um conjunto de perguntas derivado do estudo de Emil *et al.* [9], no qual participantes do Amazon Mechanical Turk produziram perguntas com listas explícitas de alternativas. Cada instância inclui a pergunta completa, as opções de resposta esperadas e uma categoria conversacional, como atividade, localização, descrição, objeto/entidade, número, pessoas ou miscelânea.

Neste trabalho, esse conjunto é reaproveitado especificamente para avaliar a capacidade dos SLMs de extrair opções de escolha em perguntas que já apresentam alternativas explícitas no enunciado. Nessa tarefa, o objetivo é produzir diretamente a lista estruturada de alternativas mencionadas no enunciado, e não apenas identificar se a pergunta contém múltiplas opções. Assim, perguntas abertas sem alternativas explícitas fazem parte apenas da tarefa de classificação de intenção, não da tarefa de extração de opções. Por exemplo, para a pergunta *"what do you want to drink water diet coke or fruit juice"*, a saída esperada é *water, diet coke, fruit juice*. A base utilizada nesta etapa contém 530 instâncias, divididas

em 424 exemplos de treinamento e 106 de teste. Diferentemente da abordagem original, baseada em regras heurísticas, vocabulários derivados de imagens e técnicas tradicionais de PLN [9], esta pesquisa avalia modelos generativos ajustados para produzir diretamente a lista estruturada de opções presentes na pergunta.

2.2. Modelos Avaliados

Os SLMs representam uma abordagem voltada à execução eficiente de modelos de linguagem em dispositivos com recursos computacionais limitados, como *smartphones*, *tablets* e sistemas embarcados. Diferentemente dos LLMs, os SLMs priorizam menor custo computacional, redução do consumo de memória e maior viabilidade de inferência local, tornando-se adequados para aplicações sensíveis à latência e à privacidade [7]. Trabalhos recentes demonstram que modelos compactos podem alcançar desempenho competitivo em tarefas de processamento de linguagem natural mesmo com poucos bilhões de parâmetros.

Neste trabalho, foram avaliados diferentes SLMs recentes voltados à inferência eficiente em tarefas de linguagem natural, incluindo modelos das famílias Phi, Llama, Qwen e SmolLM. Esses modelos foram selecionados por apresentarem arquiteturas compactas, menor custo computacional e potencial de execução em ambientes com recursos limitados. Entre eles, o Phi-3 Mini se destaca por alcançar desempenho competitivo em *benchmarks* de linguagem natural mesmo com um número reduzido de parâmetros, sendo projetado para cenários de inferência eficiente e para aplicações em dispositivos locais [8].

Além dos SLMs, foi utilizado um modelo baseado em BERT como *baseline* discriminativo na tarefa de classificação de intenção. O BERT é baseado na arquitetura *Transformer Encoder* e produz representações contextuais bidirecionais de linguagem natural, sendo amplamente utilizado em tarefas de classificação textual [11].

3. Metodologia Experimental

A Figura 1 ilustra o fluxo metodológico completo adotado para as duas tarefas distintas relacionadas à CAA abordadas neste estudo: a classificação de intenção de perguntas (utilizando o *dataset* AmazonQA) e a extração de opções de escolha em linguagem natural (utilizando o conjunto MQP).

Como detalhado no diagrama, o processo compreende etapas iniciais de pré-processamento e formatação instrucional dos dados, seguidas pelo *fine-tuning* de cinco SLMs distintos por meio de QLoRA e de quantização em 4 bits. Por fim, as predições geradas na etapa de inferência são submetidas a uma avaliação quantitativa e qualitativa, adotando-se o BERT como *baseline* discriminativo para a primeira tarefa e o Gemini 3.1 Pro como referência generativa para a segunda. Para garantir comparabilidade entre os modelos avaliados, foi mantido um protocolo padronizado de treinamento e avaliação em ambas as tarefas.

3.1. Pré-processamento dos dados

O pré-processamento dos dados visou adaptar as bases de dados originais ao treinamento supervisionado de modelos generativos. Para garantir a reprodutibilidade, as divisões entre treinamento e teste de ambas as tarefas foram fixadas com semente aleatória 42.

Para a classificação de intenção, as instâncias foram convertidas para um formato instrucional baseado no *template* Alpaca [12]. A entrada contém uma instrução textual

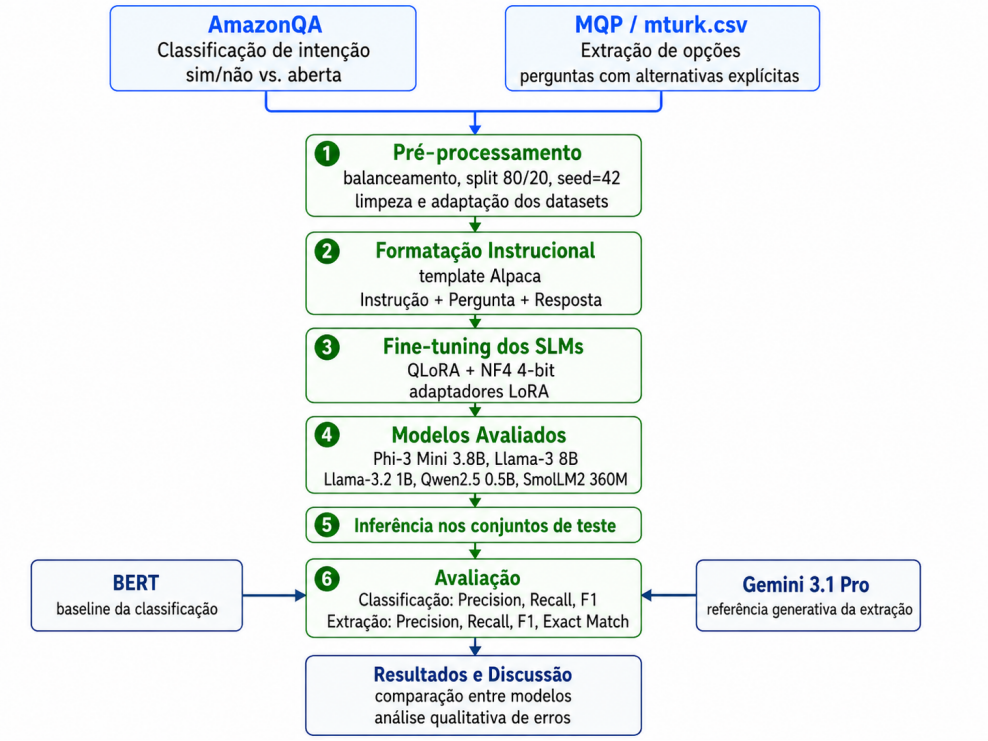

Figura 1. Visão geral do *pipeline* experimental, abrangendo desde o pré-processamento e formatação instrucional até o *fine-tuning* dos SLMs e a avaliação frente aos modelos de referência.

solicitando a classificação, seguida da pergunta original, e a saída corresponde diretamente ao rótulo da classe (*sim_nao* ou *aberta*). De maneira análoga, na extração de opções, os dados foram estruturados em formato instrucional, em que a entrada corresponde à pergunta completa e a saída esperada consiste na lista de opções extraídas, separadas por vírgula.

Durante a avaliação da extração de opções, aplicou-se uma etapa de normalização textual para reduzir penalizações decorrentes de variações superficiais nas respostas geradas. Essa normalização incluiu a remoção de artigos e conjunções em inglês (como *the, a, an, or* e *and*), além da padronização das opções para comparação direta com o gabarito.

3.2. *Fine-tuning* dos Modelos

O ajuste fino dos SLMs foi realizado via *fine-tuning* supervisionado com QLoRA [13] e quantização NF4 em 4 bits, permitindo a adaptação eficiente em uma GPU NVIDIA T4. Para reduzir o custo computacional, os pesos do modelo base foram congelados e treinaram-se apenas os adaptadores *Low-Rank Adaptation* (LoRA) [14], aplicados aos principais módulos de atenção e projeção (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj` e `down_proj`). Também utilizou-se *gradient checkpointing* para otimizar o consumo de memória. A padronização dos principais hiperparâmetros (Tabela 1) garantiu uma comparação consistente entre as diferentes escalas e famílias arquiteturais. Todos os modelos utilizaram comprimento máximo de sequência de 2048 tokens e um escalonador linear. O *batch* efetivo foi mantido constante por meio do ajuste da acumulação de gradientes em cada dispositivo. A única variação no treinamento ocorreu no número máximo

de passos (*max steps*), ajustado para acomodar a diferença de tamanho entre as bases das duas tarefas. Após a inferência, as predições geradas nos conjuntos de teste passaram por rotinas de limpeza textual para a remoção de resíduos de geração antes da avaliação.

Tabela 1. Hiperparâmetros utilizados no *fine-tuning* dos SLMs.

Parâmetro	Classificação	Extração
Quantização	NF4 4-bit	NF4 4-bit
LoRA *rank*	16	16
LoRA alpha	16	16
Batch efetivo	16	16
Max steps	500	100
Learning rate	2e-4	2e-4
Warmup steps	5	5
Otimizador	adamw_8bit	adamw_8bit
Seed	3407	3407

3.3. Métricas de Avaliação

Após o *fine-tuning*, os modelos foram avaliados nos conjuntos de teste previamente separados para cada tarefa. A avaliação foi realizada de forma independente para a classificação de intenção e para a extração de opções, considerando as diferenças entre uma tarefa discriminativa, baseada na predição de um rótulo, e uma tarefa generativa, baseada na produção de uma sequência estruturada de itens.

Na tarefa de classificação de intenção, cada modelo recebeu como entrada a pergunta do conjunto de teste e gerou como saída uma das duas classes possíveis: *sim_nao* ou *aberta*. As predições dos SLMs foram normalizadas por meio da remoção de espaços, sublinhados e resíduos textuais para permitir a comparação direta com o gabarito. Nesta tarefa, o BERT foi utilizado como *baseline* discriminativo.

Na tarefa de extração de opções, cada modelo recebeu uma pergunta com alternativas explícitas e deveria gerar como saída a lista de opções presentes na pergunta, separadas por vírgula. As predições foram comparadas ao gabarito por meio de métricas baseadas na correspondência entre os conjuntos gerados e os esperados.

O desempenho de ambas as tarefas foi avaliado utilizando-se as métricas de precisão, *recall* e F1-score, amplamente empregadas em *question answering* e extração textual [15]. A precisão mede a proporção de respostas geradas que estavam corretas; o *recall* mede a proporção de respostas esperadas que foram recuperadas pelo modelo; e o F1-score representa a média harmônica entre precisão e *recall*.

Além da avaliação quantitativa, foi realizada uma análise qualitativa de erros. Para ambas as tarefas, foram isoladas as instâncias em que as predições divergiram do gabarito, permitindo observar padrões de falha dos modelos. Na classificação de intenção, essa análise permitiu identificar perguntas ambíguas ou com uma estrutura linguística mais complexa. Na extração de opções, a análise qualitativa permitiu observar casos de omissão de alternativas, inclusão de termos extras ou falhas na delimitação correta das opções.

Por fim, o desempenho dos SLMs na classificação de intenção foi comparado ao *baseline* BERT. Já na tarefa de extração de opções, além da comparação direta entre os

modelos compactos, o Gemini foi mantido como referência de IA generativa de grande porte, para contextualizar os resultados frente a uma arquitetura mais robusta. Dessa forma, a avaliação combinou métricas quantitativas e inspeção qualitativa, permitindo analisar não apenas o desempenho agregado, mas também o comportamento dos modelos em casos específicos de erro.

4. Resultados e Discussão

Esta seção apresenta os resultados obtidos nas duas tarefas avaliadas: classificação de intenção de perguntas e extração de opções de escolha. Inicialmente, são discutidos os resultados quantitativos da classificação de intenção, comparando os SLMs ajustados com o *baseline* BERT. Em seguida, são apresentados os resultados da extração de opções, considerando a comparação entre os SLMs e a referência generativa Gemini. Por fim, são discutidos padrões qualitativos de erro e as implicações gerais dos resultados para o uso de modelos compactos em aplicações de CAA.

4.1. Classificação de Intenção

A primeira tarefa avaliada consistiu na classificação de perguntas em duas categorias: perguntas fechadas do tipo sim/não e perguntas abertas. O desempenho dos SLMs foi comparado ao *baseline* definido na metodologia. A Tabela 2 apresenta os resultados obtidos em termos de precisão, *recall* e F1-score.

Tabela 2. Resultados da tarefa de classificação de intenção.

Modelo	Precisão (%)	Recall (%)	F1-score (%)
BERT (*baseline*)	97,09	97,08	97,08
Llama-3 (8B)	96,51	96,50	96,50
Phi-3 Mini (3.8B)	94,37	94,36	94,36
Llama-3.2 (1.2B)	94,27	94,27	94,27
Qwen2.5 (0.5B)	93,68	93,68	93,68
SmolLM2 (360M)	92,54	92,52	92,52

Os resultados indicam que os SLMs ajustados via QLoRA apresentaram desempenho competitivo na tarefa de classificação de intenção. O *baseline* discriminativo BERT obteve o melhor resultado geral, alcançando 97,09% de precisão, 97,08% de *recall* e 97,08% de F1- score. Entre os modelos generativos avaliados, o Llama-3 (8B) apresentou o desempenho mais próximo ao *baseline*, atingindo 96,51% de precisão, 96,50% de *recall* e 96,50% de F1-score.

O modelo Phi-3 Mini (3.8B) também demonstrou forte capacidade de adaptação, atingindo competitivos 94,36% de F1-score. Os modelos de menor escala, como Llama-3.2-1B, Qwen2.5-0.5B e SmolLM2-360M, apresentaram desempenho ligeiramente inferior aos modelos maiores, mas ainda alcançaram resultados consistentes acima de 92% de F1-score. Esse resultado sugere que, embora arquiteturas maiores e *encoders* puros obtenham vantagem em tarefas estritamente discriminativas, modelos extremamente compactos possuem forte potencial para aplicações em cenários com maior restrição computacional.

De modo geral, os resultados da classificação de intenção demonstram que SLMs especializados podem se aproximar significativamente de um *baseline* discriminativo

forte em tarefas específicas de compreensão de perguntas. Para aplicações de CAA, esse resultado é relevante, pois sugere a viabilidade de utilizar modelos generativos compactos para identificar automaticamente o tipo de pergunta e apoiar interfaces mais adaptativas.

4.2. Extração de Opções

A segunda tarefa avaliada consistiu na extração de opções de escolha a partir de perguntas contendo alternativas explícitas. Diferentemente da classificação de intenção, essa tarefa exige que o modelo gere uma sequência estruturada de itens, separando corretamente as opções presentes na pergunta. O desempenho dos modelos foi avaliado por meio das métricas de precisão, *recall* e F1-score, após a aplicação da normalização textual descrita na metodologia. A Tabela 3 apresenta os resultados obtidos.

Tabela 3. Resultados normalizados da tarefa de extração de opções.

Modelo	Precisão (%)	Recall (%)	F1-score (%)
Phi-3 Mini (3.8B)	97,48	97,17	97,30
Gemini 3.1 Pro	96,07	95,91	95,97
Llama-3 (8B)	94,81	94,65	94,72
Llama-3.2 (1.2B)	93,16	92,47	92,74
Qwen2.5 (0.5B)	91,31	91,38	91,21
SmolLM2 (360M)	86,31	86,32	86,06

Os resultados indicam que os SLMs ajustados foram capazes de realizar a extração de opções com desempenho elevado, mesmo em uma tarefa generativa que exige a produção estruturada de múltiplos itens. O melhor desempenho foi obtido pelo Phi-3 Mini, que alcançou 97,48% de precisão, 97,17% de *recall* e 97,30% de F1-score. Esse resultado supera tanto os demais SLMs avaliados quanto a referência generativa Gemini 3.1 Pro, que obteve 96,07% de precisão, 95,91% de *recall* e 95,97% de F1-score.

O Llama-3 apresentou desempenho intermediário, com 94,72% de F1-score, enquanto o Llama-3.2 obteve 92,74%. Já os modelos de menor escala, Qwen2.5-0.5B e SmolLM2-360M, apresentaram queda mais acentuada, especialmente o SmolLM2, que obteve 86,06% de F1-score.

De forma geral, os resultados sugerem que o ajuste fino supervisionado permitiu aos SLMs aprenderem padrões relevantes para a extração estruturada de opções. O desempenho do Phi-3 Mini é particularmente relevante, pois indica que um modelo compacto e especializado pode superar uma referência generativa de maior porte em uma tarefa específica e bem delimitada. Esse resultado reforça a hipótese de que SLMs ajustados para tarefas específicas podem ser adequados para aplicações assistivas em que o interlocutor apresenta perguntas com alternativas explícitas e o sistema precisa gerar respostas estruturadas, com baixo custo computacional e potencial de execução em ambientes com recursos limitados.

4.3. Análise Qualitativa dos Erros

Na tarefa de classificação de intenção, as divergências concentraram-se principalmente em perguntas longas, ambíguas ou com contexto adicional no mesmo enunciado. Em diversos

casos, perguntas fechadas com informações contextuais foram classificadas como abertas, especialmente quando envolviam compatibilidade, atributos visuais, funcionamento de produtos ou especificações técnicas. A inspeção manual também evidenciou que parte das divergências está associada à própria adaptação do AmazonQA para uma tarefa binária, pois algumas perguntas apresentam múltiplas intenções ou rótulos discutíveis.

No conjunto de teste, foram identificadas 30 divergências no BERT, 36 no Llama-3, 58 no Phi-3 Mini, 59 no Llama-3.2, 65 no Qwen2.5-0.5B e 77 no SmolLM2-360M. Ao todo, 137 instâncias apresentaram divergência em pelo menos um dos modelos. Por esse motivo, a Tabela 4 apresenta exemplos selecionados manualmente nos quais o gabarito é mais defensável e o padrão de erro pode ser interpretado de forma consistente.

Tabela 4. Exemplos de erros qualitativos na tarefa de classificação de intenção.

Padrão observado	**Exemplo**
Contexto adicional	**Pergunta:** *There is no response to the Feb. 18, 2013 post, is this an OEM item?* **Gabarito:** *sim_nao* **Predição recorrente:** *aberta* (Llama-3, Phi-3 Mini, Llama-3.2, Qwen2.5 e SmolLM2)
Atributo visual em pergunta fechada	**Pergunta:** *The color pictured in the pics on Amazon, is that brown???* **Gabarito:** *sim_nao* **Predição recorrente:** *aberta* (Llama-3, Phi-3 Mini, Llama-3.2, Qwen2.5 e SmolLM2)
Contexto técnico adicional	**Pergunta:** *So, do you have to have a Tivo account to use this? I just need a wireless adapter for a second computer in my house.* **Gabarito:** *sim_nao* **Predição recorrente:** *aberta* (Llama-3, Llama-3.2, Qwen2.5 e SmolLM2)

Na tarefa de extração de opções, a análise qualitativa concentrou-se nos modelos de menor escala, especialmente SmolLM2-360M, Qwen2.5-0.5B e Llama-3.2-1B, por apresentarem maior interesse para cenários de restrição computacional e por evidenciarem limitações relevantes da abordagem. No SmolLM2-360M, foram identificadas 29 divergências, enquanto no Llama-3.2-1B foram observadas 16 divergências. Além disso, foram encontrados 11 erros compartilhados entre Qwen2.5- 0.5B e SmolLM2-360M, indicando casos de dificuldade comuns a modelos com menos de 1B de parâmetros.

Os erros envolveram principalmente omissão de alternativas, inclusao de termos contextuais, segmentação incorreta de entidades compostas e perda de modificadores relevantes. A Tabela 5 apresenta exemplos desses padrões, incluindo casos em que termos do contexto, como *grocery shopping*, foram extraídos indevidamente, entidades como *chase bank* e *kaiser permanente* foram fragmentadas, e termos compostos como *disney world* foram aglutinados.

Tabela 5. Exemplos de erros qualitativos na tarefa de extração de opções.

Categoria	Exemplo
Fragmentação de entidade	**Pergunta:** *of all the places you've worked did you like ibm chase bank or kaiser permanente best* **Gabarito:** *ibm, chase bank, kaiser permanente* **Predição:** *ibm, chase, bank, kaiser, permanente* (Qwen2.5/SmolLM2)
Vazamento de contexto	**Pergunta:** *do you want to go grocery shopping today or tomorrow* **Gabarito:** *today, tomorrow* **Predição:** *grocery shopping, tomorrow* (Qwen2.5)
Junção indevida	**Pergunta:** *who do you enjoy being around the most family friend partner children or stranger* **Gabarito:** *family, friend, partner, children, stranger* **Predição:** *family friend, partner, children, stranger* (Qwen2.5)
Morfologia	**Pergunta:** *do you have preteens or teens* **Gabarito:** *preteens, teens* **Predição:** *preteens, teenagers* (SmolLM2)
Aglutinação	**Pergunta:** *would you prefer a vacation to disneyland disney world or universal studios* **Gabarito:** *disneyland, disney world, universal studios* **Predição:** *disneyland, disneyworld, universalstudios* (SmolLM2)

De modo geral, a análise qualitativa mostra que os erros dos modelos não se distribuem apenas de forma aleatória, mas se concentram em estruturas linguisticamente mais ambíguas ou mais difíceis de interpretar. Na classificação de intenção, a principal dificuldade está em distinguir a intenção predominante da pergunta quando há contexto adicional, formulações pouco diretas ou possíveis ambiguidades decorrentes da adaptação do AmazonQA para uma tarefa binária. Na extração de opções, a dificuldade está em delimitar corretamente quais termos constituem alternativas válidas, especialmente quando as opções são compostas por múltiplas palavras ou aparecem misturadas ao restante do enunciado. Para aplicações de CAA, esses erros são relevantes, pois podem levar à geração de cartões incompletos, opções incorretas ou classificações inadequadas da intenção do usuário.

5. Conclusão

Este trabalho investigou o uso de SLMs em tarefas de processamento de linguagem natural relacionadas à CAA, com foco em modelos compactos e potencialmente adequados a cenários de execução com recursos computacionais limitados. Para isso, foram avaliadas duas tarefas: classificação de intenção de perguntas, distinguindo perguntas do tipo sim/não e perguntas abertas, e extração de opções de escolha a partir de perguntas contendo alternativas explícitas.

Os resultados obtidos indicam que SLMs ajustados por meio de QLoRA e quantização em 4 bits podem alcançar desempenho competitivo em tarefas específicas de compreensão e geração estruturada de linguagem. Na classificação de intenção, o *baseline* discriminativo baseado em BERT obteve o melhor resultado geral, seguido de perto pelo Llama-3. Por outro lado, na tarefa de extração de opções, o Phi-3 Mini obteve o melhor desempenho geral, superando a referência generativa Gemini 3.1 Pro nas métricas avaliadas. Esses resultados sugerem que modelos compactos, quando especializados por *fine-tuning* supervisionado, podem se aproximar de modelos discriminativos fortes em tarefas de classificação e superar referências generativas de maior porte em tarefas estruturadas e bem delimitadas.

A análise qualitativa dos erros mostrou que as principais dificuldades dos modelos estão associadas a perguntas com contexto adicional, formulações ambíguas e delimitação incorreta de alternativas. Na classificação de intenção, parte das divergências também

refletiu limitações da adaptação do AmazonQA para uma tarefa binária, especialmente em perguntas com múltiplas intenções ou rótulos discutíveis. Na extração de opções, erros como omissão de itens, inclusão de termos contextuais e fragmentação de entidades compostas são particularmente relevantes em aplicações de CAA, pois podem impactar diretamente a geração de cartões, opções de resposta ou elementos de interface apresentados ao usuário.

Apesar dos resultados promissores, este trabalho possui limitações. Os experimentos foram conduzidos em ambiente de treinamento e testes baseado em GPU T4, sem avaliação direta em dispositivos móveis ou sistemas embarcados. Além disso, a tarefa de extração de opções utilizou um conjunto de dados relativamente pequeno, foi restrita a perguntas que já apresentam alternativas explícitas no enunciado e não contempla perguntas abertas em que as possíveis respostas precisam ser geradas sem estarem explicitamente presentes na pergunta. As predições do Gemini foram obtidas via interface *web*. Outra limitação está relacionada ao uso de bases adaptadas: o AmazonQA não foi originalmente construído para cenários de CAA nem para a distinção binária proposta, o que pode introduzir ambiguidades em alguns exemplos da classificação de intenção.

Como trabalhos futuros, pretende-se avaliar os modelos em dispositivos reais, medir latência e consumo de memória em cenários de inferência local, ampliar os conjuntos de dados utilizados e investigar estratégias de integração dos modelos a interfaces reais de CAA. Também se pretende construir uma base de dados mais próxima de situações reais de comunicação assistiva, possivelmente combinando exemplos sintéticos controlados e validação por especialistas, além de investigar mecanismos de validação das saídas geradas, confirmação pelo usuário ou cuidador antes da criação de cartões, filtros para detectar opções incompletas ou inconsistentes e combinação entre SLMs e regras de segurança para reduzir classificações inadequadas da intenção do usuário.

Referências

[1] João P. C. Uchoa, Taciana Pontual Falcão, André C. Nascimento, Péricles B.C. Miranda, and Rafael Ferreira Mello. Fostering autonomy through augmentative and alternative communication. In *2021 International Conference on Advanced Learning Technologies (ICALT)*, pages 320–324. IEEE, 2021.

[2] Sarah Creer, Pamela Enderby, Simon Judge, and Alex John. Prevalence of people who could benefit from augmentative and alternative communication (aac) in the uk: determining the need. *International Journal of Language and Communication Disorders*, 51(6):639–653, 2016.

[3] Taciana Pontual Falcão, Karina Machado, Carlos Pereira, Paulo Rodrigues, Ana Paula Furtado, and André Nascimento. Eu também quero brincar! ia gerativa integrada à comunicação alternativa para inclusão de crianças neurodivergentes em contextos lúdicos. In *Anais Estendidos do XXII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, Brazil, 2023.

[4] Rodica Neamtu, André Camara, Carlos Pereira, and Rafael Ferreira. Using Artificial Intelligence for Augmentative Alternative Communication for Children with Disabilities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11746 LNCS:234–243, 2019. ISBN: 9783030293802.

[5] Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K. Kane. "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*. ACM, 2023.

[6] Fuwen Tan, Royson Lee, Łukasz Dudziak, Shell Xu Hu, Sourav Bhattacharya, Timothy Hospedales, Georgios Tzimiropoulos, and Brais Martinez. Mobilequant: Mobile-friendly quantization for on-device language models. *arXiv preprint arXiv:2408.13933*, 2024.

[7] Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

[8] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[9] Zachary Emil, Andrew Robbertz, Richard Valente, and Cole Winsor. Towards a more inclusive world: Enhanced augmentative and alternative communication for people with disabilities using ai and nlp. Major qualifying project report, Worcester Polytechnic Institute, 2020.

[10] Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. Amazonqa: A review-based question answering task. *arXiv preprint arXiv:1908.04364*, 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[12] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.