VIII Congresso Nacional de Inovação e Tecnologia - INOVA 2025
6 e 7 de Novembro de 2025
São Bento do Sul / Santa Catarina - Brasil
ISSN 2526-3145

AstroNAOta: Um Robô Social Educacional Integrado com Google Gemini

AstroNAOta: An Educational Social Robot Integrated with Google Gemini

Guilherme Cardoso*, Lucas Alexandre Zick†, Dieisson Martinelli‡, Vivian Cremer Kalempa§

RESUMO

Neste artigo se tem a apresentação do sistema AstroNAOta, um robô social educacional que integra o NAOv6 com o Google Gemini, utilizando uma arquitetura cliente-servidor desacoplada. O sistema é projetado para responder a perguntas sobre astronomia de forma lúdica e educativa, visando crianças de 6 a 10 anos. A implementação do Vosk para reconhecimento de fala local e a utilização do Gemini para geração de respostas demonstram a eficácia dessa abordagem, proporcionando interações mais naturais e educativas com os usuários. A persona do AstroNAOta foi validada através de testes qualitativos, e sua capacidade de responder a perguntas sobre astronomia de forma precisa, criativa e adequada ao público infantil. A estratégia implementada para mitigar o problema de latência no primeiro ciclo de interações mostrou-se eficaz, reduzindo significativamente o tempo de espera do usuário.

PALAVRAS-CHAVE: Robô NAO; Google Gemini; Chatbot; Robô Social; Inteligência Artificial.

ABSTRACT

This article presents the AstroNAOta system, an educational social robot that integrates the NAOv6 with Google Gemini, utilizing a decoupled client-server architecture. The system is designed to answer questions about astronomy in a playful and educational manner, targeting children aged 6 to 10 years. The implementation of Vosk for local speech recognition and the use of Gemini for response generation demonstrate the effectiveness of this approach, providing more natural and educational interactions with users. Qualitative tests validated the AstroNAOta persona, confirming its ability to answer astronomy-related questions accurately, creatively, and appropriately for the target audience. The warm-up strategy implemented to mitigate cold start issues proved effective, significantly reducing latency in the user's first interaction.

KEYWORDS: NAO Robot; Google Gemini; Chatbot; Social Robot; Artificial Intelligence.

1 INTRODUÇÃO

Os avanços na robótica social e na inteligência artificial (IA) têm impulsionado a busca por agentes capazes de interagir com humanos de forma cada vez mais natural e adaptativa. Esta evolução é crucial, especialmente em aplicações onde a comunicação expressiva e contextual é fundamental, como em ambientes educacionais e de assistência. A capacidade de um robô de mimetizar o comportamento humano, particularmente em aspectos como fala expressiva, emoção e interação contextual, é um foco central da pesquisa atual. Por exemplo, Tuttosí et al. (2025) demonstram o uso de ferramentas de *Text-to-Speech* (TTS) expressivo em robôs sociais para simular variações humanas na fala ao longo do tempo (Tuttosi et al., 2025). Além disso, Mathur et al. (2024) apontam que o desenvolvimento de agentes socialmente inteligentes exige a integração de múltiplas modalidades, incluindo percepção afetiva, cognição e linguagem, para gerar interações humanas mais

* Universidade do Estado de Santa Catarina, São Bento do Sul, Santa Catarina, Brasil. guilherme.cardoso01@edu.udesc.br.
† Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil. zick@alunos.utfpr.edu.br.
‡ Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil. dmartinelli@alunos.utfpr.edu.br.
§ Universidade do Estado de Santa Catarina, São Bento do Sul, Santa Catarina, Brasil. vivian.kalempa@udesc.br.

naturais (Mathur; Liang; Morency, 2024). Essa tendência é reforçada por Hurtado et al. (2021), que destacam a importância de comportamentos sociais aprendidos por observação humana para refletir padrões de sociabilidade em ambientes cotidianos (Hurtado; Londoño; Valada, 2021).

Apesar desses avanços, a interação por meio da linguagem falada em robôs e dispositivos digitais ainda apresenta desafios significativos. Embora a comunicação interpessoal humana seja naturalmente ajustada conforme o interlocutor e o ambiente, a maioria dos sistemas automatizados ainda não possui essa capacidade adaptativa, mantendo parâmetros fixos de fala que podem dificultar a compreensão por parte dos usuários (Ren et al., 2024). Este desafio não é novo; o primeiro *chatbot* da história, ELIZA, criado por Joseph Weizenbaum em 1966, já evidenciava como legados computacionais podiam guiar padrões de diálogo automatizado (Weizenbaum, 1966). Embora sua lógica de correspondência por padrões e substituições ainda embase inúmeras aplicações atuais, a adaptabilidade contextual da fala permanece um gargalo (Weizenbaum, 1966; Shrager, 2024). Superar essa limitação tem levado ao desenvolvimento de sistemas capazes de adaptar parâmetros vocais, como volume, taxa de fala e timbre, às condições acústicas e contextuais do ambiente, aprimorando a clareza e a naturalidade na interação com os usuários e impulsionando o desenvolvimento de modelos de IA cada vez mais complexos (Ren et al., 2024).

Neste cenário de busca por interações mais avançadas e adaptativas, a integração de modelos de linguagem de larga escala (LLMs) como o Google Gemini com robôs sociais é apontada como uma solução promissora. Imran e Almusharraf (2024) descrevem o Google Gemini como uma ferramenta de IA multimodal, capaz de processar dados de diversas fontes (texto, imagem, áudio e vídeo) e fornecer respostas avançadas, precisas e contextualmente relevantes (Imran; Almusharraf, 2024). A versatilidade do Gemini, mesmo em suas versões mais acessíveis, introduz a potência e a complexidade dos modelos de linguagem avançados no cotidiano dos usuários. Paralelamente, o robô NAO, desenvolvido pela SoftBank Robotics, destaca-se em ambientes de pesquisa e ensino por sua arquitetura modular, sistema operacional NAOqi e recursos de fala integrada, juntamente com sensores que permitem percepções avançadas do ambiente e interações multilíngues. Um exemplo emblemático é o *Project Hermes: The Socially Assistive Tour-Guiding Robot*, que utilizou um NAO6 para conduzir visitas guiadas interativas, empregando seus módulos sensoriais e motores para navegação autônoma e comunicação em tempo real (Khater; Sun; Camou, 2023).

Neste contexto, surge o AstroNAOta, um robô social educacional que visa transformar a aprendizagem de astronomia para crianças de 6 a 10 anos por meio de interações lúdicas e educativas. O AstroNAOta integra o robô NAOv6 com o Google Gemini, utilizando uma arquitetura cliente-servidor desacoplada que otimiza o desempenho e a flexibilidade do sistema. A escolha do Vosk para reconhecimento de fala local garante processamento ágil e privacidade, enquanto a utilização do Google Gemini para geração de respostas possibilita interações ricas, precisas e adequadas ao público infantil, com uma persona envolvente. Para uma melhor experiência do usuário, é destacado a estratégia de *warm-up* eficaz para o Gemini, mitigando a latência inicial e assegurando uma primeira interação rápida e natural.

Neste artigo, a arquitetura do sistema AstroNAOta e a implementação de suas funcionalidades principais são detalhadas. Onde informam-se os resultados dos testes qualitativos que validaram a persona do robô e sua capacidade de fornecer informações astronômicas de forma precisa, criativa e divertida, destaca-se a eficácia da estratégia de *warm-up* na redução da latência do sistema.

2 TRABALHOS RELACIONADOS

No contexto de integração entre *chatbots* e robôs sociais, diversos estudos têm explorado abordagens que variam em complexidade e propósito. Um exemplo notável é o trabalho de Bertacchini

et al. (2023), que como objetivo se teve a integração do robô Pepper ao sistema OpenAI (ChatGPT) para facilitar diálogos em tempo real com indivíduos portadores de Transtorno do Espectro Autista (TEA). Utilizando um modelo de linguagem avançado, o ChatGPT, os autores conseguiram proporcionar interações mais naturais e adaptadas às necessidades específicas dos usuários com TEA, demonstrando o potencial de tecnologias de IA de ponta em aplicações terapêuticas (Bertacchini et al., 2023). Como mostrado na Figura 1, o robô Pepper utiliza recursos visuais no *tablet* e gestos corporais para ilustrar emoções como felicidade e tristeza, promovendo o desenvolvimento de habilidades sociais por meio de jogos simbólicos cooperativos com a criança.

Figura 1 – Interação do robô Pepper com o usuário

Fonte: Bertacchini et al. (2023)

No trabalho de Wilcock et al. (2020), foi implementado o sistema de diálogo *WikiTalk*, que capacita robôs a conversar fluentemente sobre milhares de tópicos, utilizando como base de conhecimento a Wikipédia. O sistema foi projetado para seguir as mudanças de interesse do usuário, permitindo transições de tópico suaves durante a conversa (Wilcock; Jokinen, 2020). No caso do robô, a interação utiliza o rastreamento facial (*face-tracking*) para se voltar ao interlocutor, o que otimiza a orientação de seus microfones para uma melhor captação da fala em conversas individuais (Wilcock; Jokinen, 2020).

Em contraste com essas propostas baseadas em IA generativa e multimodalidade, o trabalho de Kuo et al. (2021) apresenta uma arquitetura significativamente mais complexa, fundamentada em redes neurais recorrentes (GRU) e fusão multissensorial de dados visuais e sonoros para fornecer respostas contextuais em robôs companheiros humanoides (Kuo et al., 2021). Embora a proposta amplie as capacidades perceptivas do robô ao integrar reconhecimento de imagens e contexto acústico, essa complexidade exige alto poder computacional e treinamento intensivo, o que pode limitar sua aplicação em cenários educacionais mais simples ou de prototipagem rápida (Kuo et al., 2021).

No desenvolvimento do AstroNAOta foi proposto fundamentalmente adotar uma arquitetura cliente-servidor desacoplada. Destacando essa abordagem para o contexto educacional, contorna a obsoloscência, externalizando o processamento itensivo, permitido que robôs com *hardware* limitado acessem a IA mais avançada disponível, como o Google Gemini, sem a necessidade de atualizações constantes no software embarcado do robô. Além disso, a solução utiliza o Vosk para reconhecimento automático de fala (ASR) local, garantindo baixa latência e maior privacidade, especialmente em

ambientes com crianças. A escolha do Vosk para ASR local foi estratégica, visando baixa latência e maior privacidade por não depender de serviços em nuvem para a transcrição, um aspecto crucial em ambientes com crianças. Assim, a arquitetura não apenas resolve um problema técnico, mas também se alinha com as boas práticas de interação humano robô (IHR), defendendo uma solução engenhosamente simples e eficaz.

3 DESENVOLVIMENTO

O desenvolvimento do sistema AstroNAOta envolveu a criação de uma arquitetura desacoplada, onde o robô NAOv6 atua como cliente e um computador com Python 3+ e GPU (se disponível) serve como o "servidor" de IA. Essa configuração permite que o robô se concentre na interação física, enquanto o processamento intensivo de IA é realizado no servidor. O servidor de IA é responsável por todo o processamento complexo, incluindo o ASR com o Vosk, que transcreve continuamente o diálogo do usuário em português. O texto resultante alimenta o "pensamento" do sistema, sendo enviado à API do Google Gemini, um LLM, cuja persona "AstroNAOta" é moldada por um *prompt* detalhado que define seu tom e estilo. A comunicação com o robô é gerenciada por uma interface construída com o *framework* Flask, que expõe uma rota específica para o NAO consultar as respostas geradas.

3.1 ARQUITETURA PROPOSTA

O sistema AstroNAOta baseia-se em uma arquitetura cliente-servidor desacoplada para maximizar a flexibilidade e superar as limitações do *hardware* do robô, onde o robô NAOv6 atua como cliente, enquanto um computador ou *notebook* com Python 3+ e GPU (se disponível) serve como "servidor" de IA, atuando como o "cérebro" do sistema. Esta configuração permite que o robô se concentre na interação física, enquanto o processamento intensivo de IA é realizado no servidor.

O servidor de IA, operando em um computador com Python 3+, é responsável por todo o processamento complexo. A "escuta" é realizada por meio de um microfone externo, com a utilização do modelo Vosk para o ASR, que transcreve continuamente o diálogo do usuário em português. O texto resultante alimenta o "pensamento" do sistema, sendo enviado à API do Google Gemini, um LLM, cuja persona "AstroNAOta" é moldada por um *prompt* detalhado que define seu tom e estilo. Por fim, a comunicação com o robô é gerenciada por uma interface construída com o *framework* Flask, que expõe uma rota específica para o NAO consultar as respostas geradas.

A escolha do Vosk para o ASR, no lugar do módulo nativo do robô (AL-SpeechRecognition), foi uma decisão técnica motivada por desafios significativos documentados em projetos anteriores. O sistema de reconhecimento de fala embarcado no NAO, embora integrado, apresenta instabilidades críticas; projetos relatam falhas na ativação do serviço, baixa confiança na transcrição e acionamento indevido por ruídos aleatórios em vez de fala humana (Khater; Sun; Camou, 2023). Adicionalmente, a documentação para implementações autônomas que não utilizam a interface gráfica Choregraphe é escassa, dificultando a depuração desses problemas (Khater; Sun; Camou, 2023).

Como alternativa, soluções baseadas em nuvem, como o Google *Speech-to-Text*, foram avaliadas. Contudo, sua integração implicaria em uma alta complexidade arquitetônica para gerenciar a incompatibilidade entre a versão de Python do robô (2.7) e os requisitos de bibliotecas modernas (3.8+) (Khater; Sun; Camou, 2023).

Tendo esse cenário específico, a escolha do Vosk se apresentou como a solução mais adequada, pois é um *kit* de ferramentas de código aberto que funciona de forma *offline* (Khater; Sun; Camou, 2023). Essa abordagem alinha-se perfeitamente com a arquitetura do AstroNAOta, delegando o

processamento de fala a um componente externo mais robusto e moderno, garantindo maior precisão e confiabilidade sem depender de conectividade com a internet para a transcrição.

Em contrapartida, o cliente robótico, executado no NAOv6 com seu ambiente Python 2.7, atua como a "voz e corpo" do sistema. Seu *script* principal é minimalista e focado em realizar requisições periódicas (*polling*) ao servidor para buscar novas respostas, sem executar qualquer processamento de linguagem. Ao receber um texto, o NAO utiliza seus módulos nativos, como o *ALTextToSpeech* para vocalização e o *ALLeds* para fornecer *feedback* visual, garantindo uma interação social, seja visual ou auditiva com o usuário, como demonstrado na Figura 2.

Figura 2 – Leds do robô NAOv6

A escolha de uma arquitetura cliente-servidor foi uma decisão primordial para contornar as limitações de *hardware* do NAOv6 e a incompatibilidade entre seu ambiente Python 2.7 e as bibliotecas modernas de ASR e LLM, que exigem Python 3+. A comunicação entre as duas partes é feita via uma API REST local, que utiliza o *framework* Flask para criar uma interface leve e padronizada. Esta natureza desacoplada é fundamental para a longevidade do projeto, pois permite que o "cérebro" do sistema evolua com os avanços em IA no servidor, sem a necessidade de modificar o *software* embarcado e robusto do robô. A decisão de usar um microfone externo para o ASR, por exemplo, foi crucial para garantir a qualidade da captura de áudio, evitando ruídos indesejados no processamento do robô. Além disso, o processamento local de fala pelo Vosk favorece a privacidade e um controle da latência, enquanto o uso de um LLM em nuvem, como o Gemini, garante respostas com qualidade de ponta.

A adoção de uma arquitetura cliente-servidor entre o robô NAOv6 e um servidor com Python 3+ reflete diretrizes recentes em *frameworks* modulares aplicados à robótica social, como demonstrado por Bono et al. (Bono et al., 2024). O processamento de fala via microfone externo, usando modelos como Vosk, evita limitações do hardware embarcado e segue diretrizes de desacoplamento funcional conforme estudos de Rikasofiadewi e Prihatmanto (Rikasofiadewi; Prihatmanto, 2016). A integração com modelos de LLM por meio de APIs RESTful está alinhada com práticas recomendadas para sistemas escaláveis baseados em LLMs (Deng et al., 2025).

A conclusão apresentada é que a interação do sistema AstroNAOta é orquestrada por uma arquitetura cliente-servidor desacoplada, projetada para dividir as responsabilidades e superar as limitações de *hardware* do robô. Neste modelo, toda a carga de processamento de IA desde ASR com o Vosk até a geração de respostas com o Google Gemini é executada no servidor. O robô NAO, por sua vez, atua exclusivamente como cliente, responsável pela interação física: consultar o servidor por novas respostas através de um ciclo de sondagem (*polling*), vocalizá-las e apresentar *feedback* visual.

O processo de interação entre o usuário, o robô NAO e o servidor de IA ocorre em estapas bem definidas, tendo como os dois principais "motores" para o funcionamento do sistema o servidor (Computador com Python 3+ sendo o "cérebro") e o cliente (Robô NAO com Python 2.7 sendo - "A Voz e Corpo"). A exemplificação visual de como esses dois fatores ocorre é apresentada na Figura 3.

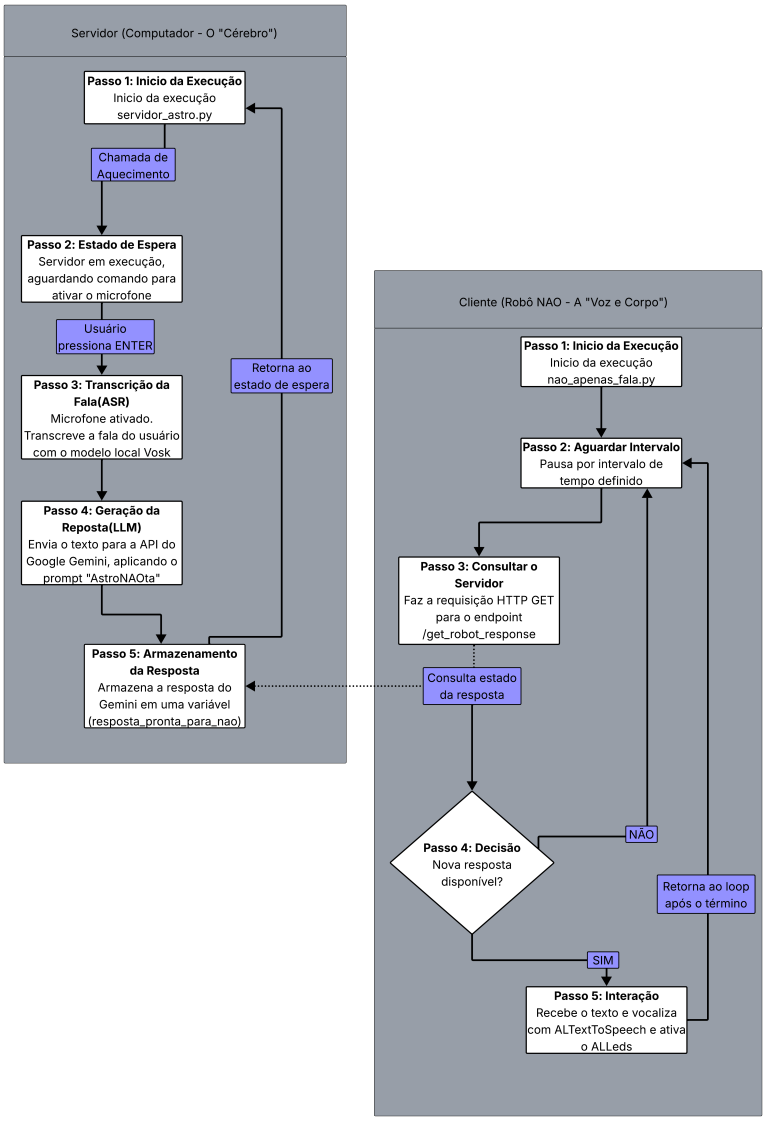

Figura 3 – Fluxograma da arquitetura cliente-servidor do sistema AstroNAOta

3.2 DESENVOLVIMENTO DA PERSONA "ASTRONAOTA" VIA ENGENHARIA DE *PROMPTS*

A materialização do sistema AstroNAOta envolveu a configuração de ambientes de desenvolvimento distintos para o servidor e para o robô NAO, além de um trabalho cuidadoso de engenharia de

prompts para dar vida à persona educativa do robô. A essência da personalidade e do conhecimento do AstroNAOta não reside em código complexo no robô, mas em um *system prompt* cuidadosamente elaborado, enviado ao LLM (Google Gemini) a cada interação. Este *prompt* atua como um "manual de instruções" para o modelo, definindo papel, tom, restrições e base de conhecimento (Chaudhary et al., 2025).

A criação do *dataset_astronomia_criancas.txt* representou um esforço de curadoria fundamental, fornecendo exemplos de interações e informações factuais adequadas para ancorar as respostas do LLM, aumentando a segurança e relevância educacional aspecto crucial ao lidar com LLMs que podem, por vezes, gerar conteúdo inadequado. Conforme destaca Liang et al. (2025), a integração de exemplos específicos (*few-shot prompting*) e instruções de domínio melhora substancialmente a estabilidade e precisão da geração em tarefas educacionais (Liang; Kalaleh; Mei, 2025).

A criação da persona AstroNAOta foi realizada por meio de princípios de engenharia de prompts, alinhados às melhores práticas da área. O *prompt* de sistema inicia com uma instrução clara e específica, definindo o papel do robô e seu público-alvo crianças de 6 a 10 anos o que orienta o LLM a modular o tom para ser entusiasmado, educativo e seguro. Foram impostas restrições de formato, como a exigência de respostas curtas e a proibição de emojis, decisões técnicas que visam facilitar a compreensão infantil, otimizar a vocalização no TTS do NAO e reduzir a latência da interação. Estratégias como *few-shot prompting* foram aplicadas com exemplos curados para promover interações seguras e contextualmente apropriadas. Segundo Asseri et al. (2025), esse cuidado é essencial na modelagem de agentes educacionais para crianças, especialmente ao tratar de emoções, vocabulário e adequação cultural (Asseri et al., 2025).

A Tabela 1 ilustra e detalha os componentes do *System Prompt* utilizado para a persona AstroNAOta, ilustrando como cada instrução contribui para o comportamento final do robô.

3.3 DESAFIOS PEDAGÓGICOS E ÉTICOS

Apesar do êxito técnico na implementação e validação da persona, o projeto AstroNAOta enfrenta desafios pedagógicos e éticos relevantes. A mediação do aprendizado por IA, embora promissora para engajamento, pode comprometer o desenvolvimento da autonomia crítica da criança, limitando sua capacidade de lidar com ambiguidades e construir perguntas próprias (Holmes et al., 2022). A ausência de um educador humano tende a restringir o aprofundamento conceitual e a correção de concepções equivocadas, apontando para a necessidade de um uso complementar da tecnologia, e não substitutivo (Luckin, 2018). Além disso, a antropomorfização de robôs sociais como o NAO pode levar ao apego excessivo, levantando questões sobre impactos no desenvolvimento socioemocional das crianças e exigindo diretrizes claras para interações saudáveis (Sharkey, 2016).

4 TESTES

Nos testes das implementações do sistema AstroNAOta, a viabilidade de integrar um robô social com um modelo de linguagem avançado como o Google Gemini foi demonstrada, utilizando uma arquitetura cliente-servidor desacoplada.

Modelos de LLM, como o Google Gemini, frequentemente enfrentam o problema conhecido como *cold start*, que ocorre quando o modelo ainda não foi carregado na memória ativa ou os recursos de GPU ainda não foram alocados. Esse fenômeno introduz uma latência significativa na primeira inferência realizada após a inicialização, afetando negativamente sistemas interativos que exigem respostas imediatas (Ghosh, 2024).

A validação da eficácia do sistema AstroNAOta foi realizado um teste qualitativo com um

Tabela 1 – Estrutura de *prompts* para o *chatbot* AstroNAOta, com exemplos e justificativas.

Tipo de Instrução	Conteúdo Específico do Prompt	Justificativa/Objetivo
Definição de Persona	"Você é o AstroNAOta, um robô guia espacial super divertido, amigável e muito inteligente! Sua missão é explicar astronomia para crianças..."	Estabelece o papel fundamental do LLM, sua identidade e objetivo principal.
Instrução de Tom/Estilo	Implícito em "super divertido, amigável, muito inteligente". Explicitado em como "entusiasmado, educativo, seguro".	Molda a maneira como o AstroNAOta se comunica, tornando a interação mais engajadora e apropriada para crianças.
Restrição de Formato de Resposta	"manter respostas curtas"	Melhora a compreensibilidade para crianças, facilita a síntese de fala pelo NAO e reduz a latência percebida.
Restrição de Conteúdo	"não usar *emojis*"	Garante que a saída seja puramente textual e adequada para vocalização pelo TTS do NAO.
Exemplo Few-Shot (Definição)	"Usuário: O que é um asteroide? AstroNAOta: Um asteroide é como uma grande pedra espacial que sobrou da formação do nosso Sistema Solar! A maioria vive num lugar chamado Cinturão de Asteroides, entre Marte e Júpiter."	Demonstra o tipo de linguagem, nível de detalhe e formato esperado para explicações de conceitos.
Exemplo Few-Shot (Analogia)	"Usuário: O que causa o dia e noite? AstroNAOta: É porque a Terra gira como um pião! Quando nosso lado está virado para o Sol, é dia! Quando ele se esconde do Sol, é noite!"	Mostra como usar analogias simples e compreensíveis para explicar fenômenos complexos a crianças.
Exemplo Few-Shot (Curiosidade)	"Usuário: Me conte uma curiosidade sobre os astronautas. AstroNAOta: Você sabia que os astronautas flutuam no espaço porque a gravidade é bem fraquinha lá em cima? Eles podem até dar cambalhotas no ar! Super divertido, né!"	Ilustra como fornecer fatos interessantes de forma lúdica e engajadora.
Instrução de Segurança (Implícita)	Persona de "guia", foco em "educativo", exemplos retirados de dataset curado.	Direciona o LLM para gerar o conteúdo apropriado e seguro para o público infantil.

conjunto de 25 perguntas, com o intuito da demonstração de sua capacidade para responder a perguntas sobre astronomia em gera, abrangendo desde questões factuais até solicitações criativas e entradas malformadas para testar a robustez do sistema.

Ao abranger múltiplas dimensões cognitivas como fluência, originalidade e coerência, a avaliação qualitativa permite detectar possíveis limitações do modelo, especialmente em contextos educacionais, nos quais respostas incorretas ou ambíguas podem impactar a experiência do usuário. Estudos como o de Zhao et al. (2024) confirmam que LLMs tendem a apresentar limitações em tarefas de criatividade original, enquanto mantêm bom desempenho em elaboração e estruturação de respostas (Zhao et al., 2025). Já Wenger e Kenett (2025) alertam que, embora os LLMs gerem respostas bem formatadas, há uma tendência à homogeneização criativa, o que justifica a inclusão de perguntas variadas no teste (Wenger; Kenett, 2025).

4.1 RESULTADOS

O robô NAOv6, com seu *hardware* limitado, conseguiu acessar as capacidades avançadas do Gemini, permitindo interações mais naturais e educativas com os usuários.

Para mitigar o problema de *cold start*, uma estratégia de *warm-up* foi implementada no

servidor de IA do sistema AstroNAOta. Essa técnica consiste em realizar uma chamada inicial ao LLM, com um *prompt* genérico, durante a fase de inicialização do servidor antes do início das interações com o usuário. Essa requisição antecipada permite o pré-carregamento dos pesos do modelo e a preparação do ambiente de inferência, reduzindo consideravelmente a latência percebida na primeira interação útil (Lou et al., 2025).

A efetividade dessa estratégia foi avaliada por meio de um teste de desempenho, cujos resultados encontram-se organizados na Tabela 2. Nela, observa-se que o cenário sem otimização apresentou uma latência inicial média de aproximadamente 79,0 segundos, típica de um *cold start*. Em contraste, com a estratégia de *warm-up*, essa latência caiu para 19,8 segundos e foi completamente mascarada antes da chegada do usuário. A análise evidencia uma redução de cerca de 75% no tempo de resposta, reforçando que a pré-inferência é uma solução eficaz e alinhada às melhores práticas contemporâneas em sistemas baseados em LLMs (Lou et al., 2025; Ghosh, 2024).

Tabela 2 – Comparação de Latência da Primeira Chamada à API com e sem a Estratégia de *Warm-Up*.

Métrica	Com *Warm-Up* (Na Inicialização)	Sem *Warm-Up* (Na 1ª Interação)
Latência (Duração)	~19.8 segundos	~79.0 segundos
Impacto na Experiência do Usuário	Nulo. A latência é mascarada de forma transparente durante a inicialização do servidor, antes da interação do usuário.	Médio. O usuário aguarda a latência completa, resultando em um atraso significativo na primeira resposta do robô.

A validação de persona foi observada atráves do teste qualitativo, o sistema demonstrou consistência em manter um tom lúdico e utilizar analogias simples, em uma pergunta sobre constelações, a resposta acabou sendo: "Constelações são grupos de estrelas que formam desenhos imaginários no céu! As pessoas antigamente ligavam as estrelas para criar figuras, como ursos, leões e outras coisas divertidas! É como brincar de ligar os pontos no céu." Este resultado confirma o sucesso da engenharia de *prompt* na modelagem do comportamento do AstroNAOta.

A precisão e validade dos fatos foi verificada em diversas questões, como ao identificar corretamente "Ganimedes" como a maior lua de Júpiter e contextualizá-la como sendo maior que o planeta Mercúrio, tendo em vista que a resposta completa foi: "A maior lua de Júpiter é Ganimedes! Ela é tão grande que até maior que o planeta Mercúrio! Imagine uma lua gigante orbitando um planeta gigante, que incrível!". Isso atesta a capacidade do sistema de fornecer informações corretas de forma adequada ao público infantil.

O sistema também exibiu notável capacidade criativa, indo além de respostas factuais para gerar conteúdo original. Ao ser solicitado a "contar uma lenda", o AstroNAOta elaborou uma narrativa sobre as estrelas serem "pequenas luzes deixadas pelos sonhos das crianças", mostrando a flexibilidade do LLM integrado para tarefas abertas, tendo a reposta completa sendo: "Diz a lenda que as estrelas são pequenas luzes deixadas pelos sonhos das crianças. Toda vez que uma criança tem um sonho muito bonito, uma estrelinha brilha no céu, contando sua história para o universo. Por isso, é importante sonhar muito, para que o céu fique sempre cheio de estrelas brilhantes!"

A robustez do sistema foi testada na categoria de Tratamento de Erro. Diante de uma entrada com erro de reconhecimento de fala por parte do Vosk, o sistema não apenas evitou uma falha, mas inferiu a intenção do usuário, sugerindo a correção de "Albert <ruído>" para "Albert Einstein". Esta capacidade de gerenciar imperfeições na comunicação é crucial para uma interação humano-robô eficaz, sendo a resposta completa: "Você quis dizer Albert Einstein? Ele foi um cientista muito importante que fez descobertas incríveis sobre o universo, como a famosa teoria da relatividade! Ele era super inteligente e mudou a forma como entendemos o espaço e o tempo." Em conjunto, a análise qualitativa demonstra que o sistema AstroNAOta cumpre seus objetivos, respondendo de

forma precisa, criativa e robusta, sempre alinhado à sua persona educativa.

5 CONCLUSÃO

O sistema AstroNAOta demonstrou com sucesso a viabilidade de uma arquitetura cliente-servidor desacoplada para modernizar plataformas de robótica social, como o NAOv6, superando suas limitações de *hardware*. Ao delegar o processamento intensivo de IA para um servidor externo, foi possível integrar de forma robusta o modelo Google Gemini, permitindo interações complexas e ricas que seriam impossíveis de executar no *hardware* embarcado do robô.

Os resultados confirmam duas contribuições centrais. Primeiro, a análise qualitativa da persona "AstroNAOta" atestou que a combinação do reconhecimento de fala local via Vosk com a geração de texto por meio de engenharia de *prompts* resultou em respostas precisas, criativas e contextualmente adequadas ao público infantil. Segundo, e de grande importância prática, a implementação de uma estratégia de *warm-up* mostrou-se uma solução eficaz para mitigar o problema de *cold start* do LLM, reduzindo a latência na primeira interação em aproximadamente 75%, um ganho crucial para a experiência do usuário.

Portanto, não apenas se tem a apresentação de um agente educacional interativo, mas contribui para o campo com um modelo arquitetônico replicável. Ele evidencia um caminho prático e de baixo custo para estender a vida útil de *hardware* legado, democratizando o acesso a tecnologias de IA de ponta e viabilizando novas aplicações em robótica social e educacional.

5.1 TRABALHOS FUTUROS

Para trabalhos futuros, propõe-se a ampliação do escopo do AstroNAOta para incluir uma base de conhecimento mais abrangente, incorporando temas além de astronomia. A utilização de *fine-tuning* para personalizar ainda mais o comportamento do LLM, adaptando-o a contextos não só educacionais, mas para o âmbito social.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Disponibilidade de código

O código da aplicação principal desenvolvida e discutida neste estudo está disponível no repositório GitHub: `https://github.com/Laboratorio-de-Informatica-Industrial/astroNAOta-chatbot`.

Conflito de interesse

Não há conflito de interesse .

REFERÊNCIAS

ASSERI, Bushra et al. **Deciphering Emotions in Children Storybooks: A Comparative Analysis of Multimodal LLMs in Educational Applications**. [*S. l.: s. n.*], 2025. arXiv: 2506.18201 [cs.CL]. Disponível em: https://arxiv.org/abs/2506.18201.

BERTACCHINI, Francesca et al. A social robot connected with chatGPT to improve cognitive functioning in ASD subjects. **Frontiers in Psychology**, Frontiers, v. 14, p. 1232177, 2023.

BONO, Antonio et al. **Open Access NAO (OAN): a ROS2-based software framework for HRI applications with the NAO robot**. [*S. l.: s. n.*], 2024. arXiv: 2403.13960 [cs.RO]. Disponível em: https://arxiv.org/abs/2403.13960.

CHAUDHARY, Nishi et al. Prompt to Protection: A Comparative Study of Multimodal LLMs in Construction Hazard Recognition. **arXiv preprint arXiv:2506.07436**, 2025.

DENG, Sida et al. LRASGen: LLM-based RESTful API Specification Generation. **arXiv preprint arXiv:2504.16833**, 2025.

GHOSH, Himel. Enabling Efficient Serverless Inference Serving for LLM (Large Language Model) in the Cloud. **arXiv preprint arXiv:2411.15664**, 2024.

HOLMES, Wayne et al. Ethics of AI in education: Towards a community-wide framework. **International Journal of Artificial Intelligence in Education**, Springer, v. 32, n. 3, p. 504–526, 2022.

HURTADO, Juana Valeria; LONDOÑO, Laura; VALADA, Abhinav. From learning to relearning: A framework for diminishing bias in social robot navigation. **Frontiers in Robotics and AI**, Frontiers Media SA, v. 8, p. 650325, 2021.

IMRAN, Muhammad; ALMUSHARRAF, Norah. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. **Smart Learning Environments**, Springer, v. 11, n. 1, p. 22, 2024.

KHATER, Aly; SUN, Justin; CAMOU, Fernando. Project Hermes: The Socially Assistive Tour-Guiding Robot. Santa Clara: Santa Clara University, 2023., 2023.

KUO, Ping-Huan et al. Multi-sensor context-aware based chatbot model: An application of humanoid companion robot. **Sensors**, MDPI, v. 21, n. 15, p. 5132, 2021.

LIANG, Haoran; KALALEH, Mohammad Talebi; MEI, Qipei. Integrating Large Language Models for Automated Structural Analysis. **arXiv preprint arXiv:2504.09754**, 2025.

LOU, Chiheng et al. Towards Swift Serverless LLM Cold Starts with ParaServe. **arXiv preprint arXiv:2502.15524**, 2025.

LUCKIN, Rosemary. **Machine Learning and Human Intelligence. The future of education for the 21st century**. [S. l.]: UCL institute of education press, 2018.

MATHUR, Leena; LIANG, Paul Pu; MORENCY, Louis-Philippe. Advancing Social intelligence in AI agents: technical challenges and open questions. **arXiv preprint arXiv:2404.11023**, 2024.

REN, Qiaoqiao et al. No More Mumbles: Enhancing Robot Intelligibility Through Speech Adaptation. **IEEE Robotics and Automation Letters**, IEEE, 2024.

RIKASOFIADEWI, Putri Nhirun; PRIHATMANTO, Ary Setijadi. Design and implementation of audio communication system for social-humanoid robot Lumen as an exhibition guide in Electrical Engineering Days 2015. **arXiv preprint arXiv:1607.04765**, 2016.

SHARKEY, Amanda JC. Should we welcome robot teachers? **Ethics and Information Technology**, Springer, v. 18, n. 4, p. 283–297, 2016.

SHRAGER, Jeff. ELIZA Reinterpreted: The world's first chatbot was not intended as a chatbot at all. **arXiv preprint arXiv:2406.17650**, 2024.

TUTTOSI, Paige et al. EmojiVoice: Towards long-term controllable expressivity in robot speech. **arXiv preprint arXiv:2506.15085**, 2025.

WEIZENBAUM, Joseph. ELIZA—a computer program for the study of natural language communication between man and machine. **Communications of the ACM**, ACM New York, NY, USA, v. 9, n. 1, p. 36–45, 1966.

WENGER, Emily; KENETT, Yoed. We're Different, We're the Same: Creative Homogeneity Across LLMs. **arXiv preprint arXiv:2501.19361**, 2025.

WILCOCK, Graham; JOKINEN, Kristiina. WikiTalk and WikiListen: Towards listening robots that can join in conversations with topically relevant contributions. In*:* ECAI 2020. [S. l.]: IOS Press, 2020. p. 2943–2944.

ZHAO, Yunpu et al. Assessing and understanding creativity in large language models. **Machine Intelligence Research**, Springer, v. 22, n. 3, p. 417–436, 2025.