



**FACULDADE DE TECNOLOGIA, CIÊNCIAS E EDUCAÇÃO**  
**Graduação**

**GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO**

Aplicação de Suporte de Voz em Chatbots de Inteligência  
Artificial  
para o Setor Bancário

Vitor Rian Rodrigues Paulino  
Bárbara Caroline Benato

## Resumo

A transformação digital acelerada do setor bancário evidenciou a necessidade de soluções de atendimento mais humanizadas e eficientes. Este trabalho investiga a viabilidade técnica e funcional da aplicação de Inteligência Artificial Generativa, combinada com recursos de conversão de voz (Text-to-Speech e Speech-to-Text), como ferramenta de modernização do relacionamento com o cliente. A metodologia consistiu em pesquisa bibliográfica exploratória e um estudo de caso prático, materializado no desenvolvimento de um protótipo funcional que integra o modelo multimodal Google Gemini em uma arquitetura de microsserviços. Os testes realizados demonstraram que o agente desenvolvido é capaz de manter o contexto conversacional, processar documentos e áudios simultaneamente e personalizar ofertas financeiras. Os resultados confirmam a viabilidade da solução, indicando que agentes de IA podem oferecer disponibilidade ininterrupta e experiências personalizadas, respeitando os requisitos de privacidade e ética fundamentais ao setor.

**Palavras-chave:** Inteligência Artificial; Chatbots; Atendimento Bancário; Síntese de Fala; Reconhecimento de Fala; Grandes Modelos de Linguagem; Geração Aumentada por Recuperação.

## Abstract

The accelerated digital transformation of the banking sector has highlighted the need for more humanized and efficient customer service solutions. This work investigates the technical and functional feasibility of applying Generative Artificial Intelligence, combined with voice conversion resources (Text-to-Speech and Speech-to-Text), as a tool for modernizing customer relationships. The methodology consisted of exploratory bibliographic research and a practical case study, materialized in the development of a functional prototype that integrates the Google Gemini multimodal model into a microservices architecture. The tests demonstrated that the developed agent is capable of maintaining conversational context, processing documents and audio simultaneously, and personalizing financial offers. The results confirm the solution's feasibility, indicating that AI agents can offer uninterrupted availability and personalized experiences while observing fundamental privacy and ethical requirements.

**Keywords:** Artificial Intelligence; Chatbots; Banking Service; Text-to-Speech; Speech-to-Text; Large Language Models; RAG.

## 1 Introdução

A transformação digital do setor bancário tem impulsionado a busca por soluções tecnológicas que unam eficiência operacional e qualidade no atendimento ao cliente. Embora o Brasil destaque-se globalmente pela adoção de inovações como o *Pix* e o desenvolvimento do *Drex (Real Digital)*, as instituições financeiras ainda enfrentam o desafio de transpor a personalização do atendimento humano para o ambiente digital.

Nesse contexto, a ascensão dos Grandes Modelos de Linguagem (*Large Language Models* - LLMs), como o *Google Gemini* (Google Gemini Team, 2023) e o GPT, introduziu novas possibilidades para a interação homem-máquina. Esses modelos operam processando linguagem natural e respondendo a instruções específicas, tecnicamente denominadas *prompts*. Um *prompt* consiste na entrada de dados ou comando textual fornecido à IA para guiar sua resposta, permitindo que o modelo adapte seu tom, estilo e conteúdo às necessidades do usuário.

A evolução dessas tecnologias viabilizou o desenvolvimento dos chamados *Agentes de "Inteligência Artificial" (AI Agents)*. Diferentemente de um chatbot tradicional baseado em regras rígidas, um Agente de IA é um sistema autônomo capaz de perceber o ambiente, raciocinar sobre as intenções do usuário e executar ações para atingir um objetivo específico (RUSSELL; NORVIG, 2021). Instituições como o Bradesco, com a BIA (Banco Bradesco S.A., 2023), e o *Bank of America*, com a Erica (Bank of America, 2024), já exploram esses recursos, mas a naturalidade da interação permanece uma barreira técnica.

A principal dificuldade encontrada no cenário atual reside na natureza "robótica" e impessoal dos atendimentos automatizados, que frequentemente frustram os usuários e limitam a adesão aos canais digitais. A tentativa de humanizar essa experiência depende não apenas da inteligência do texto gerado, mas da modalidade de interação. Para superar essa barreira, destacam-se as tecnologias de processamento de voz: *Speech-to-Text* (STT) e *Text-to-Speech* (TTS).

Este trabalho de conclusão de curso busca analisar a viabilidade da orquestração dessas tecnologias na mitigação da artificialidade dos sistemas atuais para uma melhor experiência do usuário.

## 2 Referencial teórico

### 2.1 Agentes de Inteligência Artificial: Da Teoria Clássica à Arquitetura de LLMs

Na computação autônoma, um agente é definido como uma entidade capaz de perceber seu ambiente por meio de sensores e exercer influência sobre o estado do mundo através de atuadores. Segundo a fundamentação clássica de Russell e Norvig (2021), a racionalidade técnica de tais sistemas reside na seleção de ações que maximizem sua utilidade esperada (*expected utility*), baseando-se em sua sequência de percepções e conhecimento prévio. Tradicionalmente, esses agentes operam sob lógicas reativas, baseadas em modelos internos ou objetivos específicos. A evolução contemporânea, impulsionada pelos Grandes Modelos de Linguagem (*Large Language Models*), transpõe esse paradigma ao utilizar o modelo como motor de raciocínio (*reasoning engine*) centralizado (Google Gemini Team, 2023).

De acordo com a arquitetura referencial proposta pela NVIDIA Varshney (2023), a construção de um agente moderno fundamenta-se em quatro módulos interdependentes: o *Agent Core*, que atua como o cérebro decisório; a Memória, subdividida em curto prazo para o fluxo de pensamento e longo prazo para armazenamento de interações; o Planejamento, que viabiliza a decomposição de tarefas complexas; e o conjunto de Ferramentas (*tool use*) para execução de fluxos de trabalho via APIs externas. Diferente de sistemas conversacionais estáticos, estes agentes manifestam agência executiva por meio de mecanismos de *Function Calling*, traduzindo intenções expressas em linguagem natural em chamadas procedurais estruturadas (Microsoft, 2024). No setor bancário, essa orquestração permite interações autônomas ancoradas em dados factuais via *Retrieval-Augmented Generation* (RAG), garantindo conformidade regulatória, rastreabilidade da informação e alta resolutividade operacional (LEWIS et al., 2020).

#### 2.1.1 Arquitetura Moderna: O Papel dos LLMs e Transformers

A transição de sistemas baseados em regras para agentes generalistas fundamenta-se na arquitetura Transformer, introduzida em “Attention is All You Need”. Este paradigma superou as limitações sequenciais das Redes Neurais Recorrentes (RNNs) através do mecanismo de *Self-Attention*, que permite a modelagem de dependências de longo prazo e o processamento paralelo massivo.

Essa base tecnológica viabilizou os *Large Language Models* (LLMs), que hoje atuam como o “cérebro” (*reasoning engine*) dos agentes modernos. Neste arranjo, o LLM processa o contexto conversacional como um estado interno dinâmico e utiliza sua capacidade de generalização para planejar ações complexas, efetivamente operando como um Agente Baseado em Utilidade capaz de lidar com a ambiguidade da linguagem

natural.

### 2.1.2 O Agente no Domínio Bancário: Agência Executiva e Tool Use

No setor financeiro, a aplicação de IA transcende a geração de texto, focando na agência executiva via *Tool Use*. Diferente de modelos puramente generativos, estes agentes possuem autonomia para orquestrar chamadas de função (*Function Calling*) e interagir com sistemas externos. A arquitetura técnica apoia-se em três pilares:

**Integração Determinística:** O modelo traduz intenções do usuário em chamadas de API estruturadas (JSON/XML) para execução segura de transações (e.g., transferências, consultas), garantindo a integridade dos dados.

**Retrieval-Augmented Generation (RAG):** Para mitigar alucinações e garantir conformidade regulatória, o agente utiliza mecanismos de recuperação vetorial para acessar bases de conhecimento internas em tempo real, fundamentando (*grounding*) suas respostas em dados factuais.

**Multimodalidade Nativa:** Adoção de modelos capazes de processar texto, áudio e vídeo em um único espaço latente (ex: Google Gemini), eliminando a latência de pipelines modulares (STT/TTS) e proporcionando interações humanizadas.

Implementações em larga escala, como a Erica (*Bank of America*) (Bank of America, 2024) e a BIA (Bradesco) (Banco Bradesco S.A., 2023), validam este modelo. Tais sistemas exemplificam a aplicação de *Trusted AI* voltada à eficiência operacional, onde o agente atua não apenas como interface conversacional, mas como um operador autônomo supervisionado.

## 2.2 O Futuro das Soluções em Atendimento de IA: Large Language Models (LLMs)

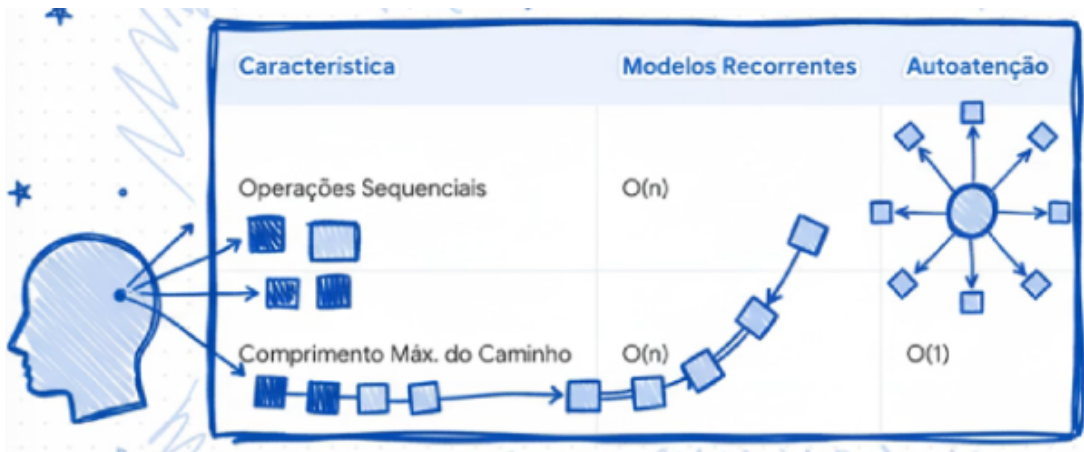
A publicação do artigo "Attention is All You Need" (VASWANI et al., 2017) estabeleceu um novo paradigma computacional ao superar as limitações das redes recorrentes e convolucionais. A arquitetura proposta solucionou a dificuldade de manutenção de contexto em longas sequências e otimizou custos de treinamento. Essa base tecnológica viabilizou o surgimento de Large Language Models (LLMs), como o ChatGPT (Microsoft, 2024), e impulsionou a evolução para padrões multimodais. Modelos recentes, como o Google Gemini (Google Gemini Team, 2023), expandem essa capacidade para a interpretação e geração simultânea de texto, imagem, áudio e vídeo, definindo a atual fronteira da IA generativa.

### 2.2.1 Transformers e a Revolução da Atenção

A arquitetura *Transformers* estabeleceu um novo padrão no processamento de linguagem natural ao substituir a recorrência sequencial pelo mecanismo de autoatenção (*Self-Attention*). Enquanto as Redes Neurais Recorrentes (RNNs) processam dados de forma linear, limitando a retenção de contexto, os Transformers analisam a sequência completa simultaneamente.

Conforme demonstrado na Figura 1, essa abordagem reduz a complexidade do caminho de conexão entre elementos distantes de linear  $O(n)$  para constante  $O(1)$  (VASWANI et al., 2017). Essa otimização viabiliza a captura eficiente de dependências de longo prazo e a paralelização massiva do treinamento, superando as limitações de escalabilidade dos modelos anteriores.

Figura 1 – Representação da diferença entre algoritmos recorrentes e de Self Attention



Fonte: Gerado pelo autor via NotebookLM 2025.

### 2.2.2 A Engenharia do Mecanismo de Atenção

O mecanismo de atenção do Transformer pode ser descrito como uma função matemática que opera com consultas (*Query*), chaves (*Key*) e valores (*Value*). Cada palavra de entrada é avaliada em relação às chaves correspondentes, gerando uma saída ponderada que atribui diferentes pesos às palavras, conforme uma função de compatibilidade.

Entre os tipos de atenção utilizados, destaca-se a *Scaled Dot-Product Attention*, que realiza a vetorização das consultas, chaves e valores, calcula o produto escalar entre consultas e chaves, divide o resultado pela raiz quadrada da dimensão das chaves  $\sqrt{D_k}$  e aplica a função *Softmax* para determinar os pesos finais.

Figura 2 – Função Scaled Dot-Product

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Fonte: Vaswani et al. (2017).

Nesta formulação, cada variável desempenha um papel específico na álgebra do modelo:

- Q (Queries/Consultas): Representa a matriz de consultas, contendo os vetores que procuram informações relevantes na sequência de entrada;
- K (Keys/Chaves): Representa a matriz de chaves, que atuam como identificadores ou rótulos para os dados disponíveis;
- V (Values/Valores): Representa a matriz de valores, que contém a informação semântica real a ser extraída;
- $d_k$ : Refere-se à dimensão dos vetores de chaves.

O processo inicia-se com o cálculo do produto escalar entre as consultas ( $Q$ ) e as chaves transpostas  $K^T$ . O resultado é dividido pela raiz quadrada da dimensão das chaves  $\sqrt{d_k}$ , um fator de escalonamento essencial. Sem este ajuste, para grandes valores de  $d_k$ , os produtos escalares poderiam atingir magnitudes elevadas, empurrando a função Softmax para regiões onde os gradientes são extremamente pequenos, o que prejudicaria o treino da rede neural (VASWANI et al., 2017).

A função Softmax é uma função de ativação que normaliza um vetor de valores reais arbitrários ( $z$ ) numa distribuição de probabilidade, onde a soma de todos os elementos é igual a 1. A sua definição matemática é dada por:

Figura 3 – Detalhamento da função softmax

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (2.1)$$

Fonte: Vaswani et al. (2017).

Onde cada componente da equação é definido conforme segue:

- $\sigma(\vec{z})_i$ : Representa o valor de saída da função para o componente  $i$  do vetor, resultando em uma probabilidade entre 0 e 1;

- $e$ : Refere-se à constante de Euler (base do logaritmo natural, aproximadamente 2,718), utilizada para a exponenciação que garante valores positivos;
- $z_i$ : Representa o  $i$ -ésimo elemento do vetor de entrada  $\vec{z}$  (score original antes da normalização);
- $K$ : Indica o número total de elementos no vetor de entrada (a dimensão do vetor sobre o qual a Softmax é aplicada);
- $j$ : Atua como o índice de iteração do somatório no denominador, percorrendo todas as posições do vetor de 1 até  $K$ ;
- $\sum_{j=1}^K e^{z_j}$ : Representa o somatório das exponenciais de todos os elementos do vetor, funcionando como o fator de normalização para que a soma final das probabilidades seja igual a 1.

Dessa forma, a Softmax garante que os pesos de atenção sejam positivos e normalizados, permitindo que o modelo "foque" em partes específicas da entrada ao multiplicar estes pesos pela matriz de valores ( $V$ ).

### 2.2.2.1 Componentes da Arquitetura: Multi-Head Attention, Encoder e Decoder

A arquitetura baseia-se fundamentalmente no mecanismo de *Multi-Head Attention*, uma extensão do *Self-Attention* que permite ao modelo processar informações de diferentes subespaços de representação simultaneamente. Neste processo, projeções lineares distintas de consultas (*queries*), chaves (*keys*) e valores (*values*) operam em paralelo, sendo seus resultados concatenados para compor a saída final, o que confere ao modelo a capacidade de capturar dependências complexas sob múltiplas perspectivas dentro de uma mesma sequência.

Estruturalmente, o *Encoder* é constituído por uma pilha de  $N$  camadas idênticas, cada uma subdividida em duas etapas principais: o mecanismo de *Multi-Head Self-Attention* e uma rede *Position-wise Feed-Forward*. Conforme estabelecido por Vaswani et al. (2017), para viabilizar as conexões residuais, todas as subcamadas e *embeddings* produzem saídas com dimensão fixa  $d_{model} = 512$ . Em contrapartida, a subcamada *Feed-Forward* opera com uma dimensionalidade interna expandida de  $d_{ff} = 2048$ , permitindo o processamento robusto de características não lineares antes da projeção de retorno ao espaço original.

O *Decoder*, por sua vez, também é composto por uma pilha de  $N$  camadas idênticas, mas diferencia-se pela integração de uma terceira subcamada de atenção que incide sobre a saída do *encoder*. Uma característica crítica deste componente

é o uso do *Masked Multi-Head Attention*. Esse mascaramento impõe uma restrição auto-regressiva essencial para tarefas de geração sequencial, assegurando que a predição para a posição  $i$  dependa estritamente das posições antecedentes ( $< i$ ). Tal mecanismo impede o vazamento de informações futuras durante o treinamento, garantindo a integridade do processo preditivo (VASWANI et al., 2017).

### 2.2.3 Retrieval-Augmented Generation (RAG)

O *Retrieval-Augmented Generation* (RAG) é uma arquitetura híbrida proposta por Lewis et al. (2020) que combina a memória paramétrica pré-treinada de um LLM com uma memória não-paramétrica externa (geralmente um índice vetorial denso). Essa abordagem soluciona duas limitações críticas dos modelos generativos padrão: a propensão a alucinações factuais e a obsolescência do conhecimento pós-treinamento.

Tecnicamente, o RAG opera através da integração de dois componentes distintos: um recuperador (*retriever*) e um gerador (*generator*). O fluxo de execução segue três etapas determinísticas:

**Recuperação (*Retrieval*):** A consulta do usuário  $x$  é convertida em uma representação vetorial (*embedding*). O sistema executa uma busca por similaridade semântica (geralmente via *Nearest Neighbor Search*) em uma base de dados externa, retornando os  $k$  documentos  $z$  mais relevantes (LEWIS et al., 2020).

**Aumentação (*Augmentation*):** Os fragmentos recuperados são concatenados à consulta original dentro da janela de contexto do modelo, formando um *prompt* enriquecido que instrui o LLM a utilizar aquelas informações específicas como fonte de verdade.

**Geração (*Generation*):** O modelo sequencial gera a resposta final  $Y$  condicionada tanto à entrada  $x$  quanto aos documentos recuperados  $z$ , maximizando a probabilidade  $p(Y|x, z)$ .

No contexto de sistemas bancários, a implementação do RAG é estratégica para garantir a conformidade (*compliance*) e a auditabilidade. Ao ancorar as respostas do agente em documentos oficiais (manuais de produtos e regulamentações), o sistema mitiga a geração de informações incorretas e permite a rastreabilidade da fonte da informação, eliminando a necessidade de *fine-tuning* frequente para atualização de regras de negócio (LEWIS et al., 2020; RUSSELL; NORVIG, 2021).

## 2.3 A Nova Composição das Inteligências Artificiais: MultiModal

A evolução recente da Inteligência Artificial consolidou o conceito de modelos multimodais, que transcendem o processamento exclusivo de texto (*text-to-text*) para

interpretar e gerar simultaneamente imagens, áudio, vídeo e código. O modelo Gemini, desenvolvido pelo Google, exemplifica o estado da arte nessa tecnologia por ser estruturado como "nativamente multimodal", treinada desde o início com diferentes tipos de dados integrados.

Conforme detalhado no relatório técnico (Google Gemini Team, 2023), a arquitetura é dividida em três variantes otimizadas para diferentes escalas de processamento, descritas pelos autores como:

“Ultra for highly-complex tasks, Pro for enhanced performance and deployability at scale, and Nano for on-device applications.” (Google Gemini Team, 2023, p. 2).

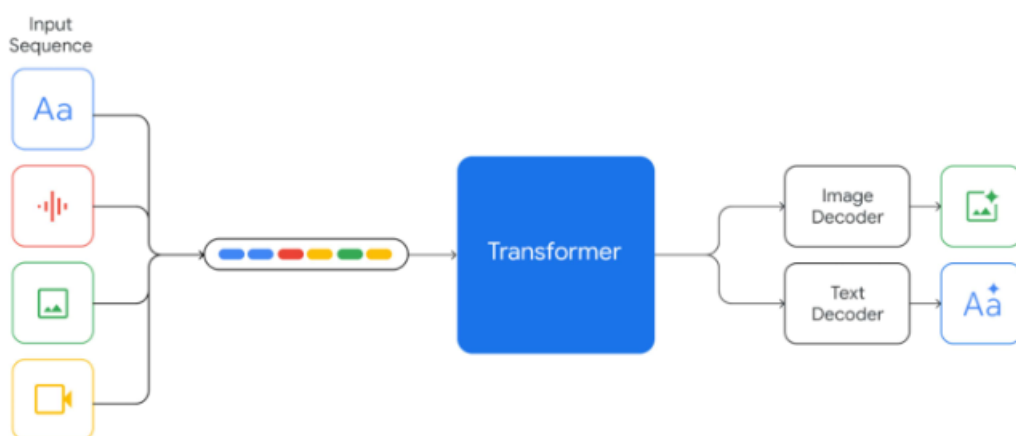
Essa segmentação permite que a arquitetura multimodal atenda desde tarefas de raciocínio complexo em *data centers* até a execução eficiente em dispositivos móveis, mantendo a coerência na interpretação de dados mistos.

### 2.3.1 Arquitetura do modelo Gemini

A arquitetura do Gemini fundamenta-se na estrutura Transformer (VASWANI et al., 2017), otimizada para o processamento multimodal nativo. Segundo o Google Gemini Team (2023), o modelo utiliza decodificadores (*decoders*) eficientes treinados em TPUs v4 e v5e, suportando sequências de entrada mistas e um contexto de até 32.000 *tokens*.

Diferentemente de modelos adaptados, o Gemini foi treinado desde a base para processar texto, imagem, áudio e vídeo de forma conjunta e intercalada. A Figura 4 ilustra esse fluxo unificado, onde sinais de áudio (16kHz) e visuais são ingeridos diretamente como *tokens* na rede, preservando a fidelidade dos dados sem a necessidade de pré-processamento externo (Google Gemini Team, 2023).

Figura 4 – Arquitetura do processamento multimodal do modelo Gemini



Fonte: Google Gemini Team (2023)

## 2.3.2 Multimodalidade expandida

A arquitetura nativamente multimodal das novas gerações de Inteligência Artificial permite o processamento integrado e simultâneo de diferentes tipologias de dados. Para compreender a extensão dessas capacidades, é fundamental analisar as evidências técnicas de desempenho apresentadas na literatura especializada. O modelo Gemini, conforme detalhado no *paper* técnico *Gemini: A Family of Highly Capable Multimodal Models* (Google Gemini Team, 2023), exemplifica o estado da arte nessa tecnologia. As subseções a seguir detalham as nuances e as métricas de validação (*benchmarks*) do modelo em suas principais modalidades de atuação.

### 2.3.2.1 Compreensão multimodal em Texto (*Text Reasoning*)

A proficiência do modelo Gemini em tarefas baseadas em texto fundamenta-se em sua capacidade de raciocínio avançado e compreensão de linguagem natural em larga escala. Conforme reportado pelo Google Gemini Team (2023, p. 5), o modelo Ultra atingiu uma pontuação de 90,04% no *benchmark* MMLU (*Massive Multitask Language Understanding*), que avalia o conhecimento em 57 disciplinas distintas, variando de matemática elementar a ética e direito profissional.

Este resultado representa um marco técnico significativo, pois o Gemini foi o primeiro modelo a superar a performance de especialistas humanos no mesmo teste, estimada em 89,8% (Google Gemini Team, 2023). Além do conhecimento enciclopédico, o modelo demonstrou desempenho superior em testes de raciocínio lógico e matemático, superando concorrentes diretos em métricas de resolução de problemas complexos, como o *Big-Bench Hard*.

No setor financeiro, essa capacidade de raciocínio textual é crítica para a viabilidade de agentes autônomos. Ela permite que sistemas automatizados não apenas recuperem informações, mas interpretem cláusulas contratuais, realizem cálculos baseados em regras descritas em linguagem natural e forneçam consultoria personalizada com nível de precisão comparável ou superior ao de analistas humanos.

### 2.3.2.2 Compreensão multimodal em Imagens (*Image Understanding*)

A capacidade de interpretação visual (*Image Understanding*) do modelo Gemini transcende a simples classificação de objetos, estendendo-se à extração estruturada de informações a partir de documentos complexos e gráficos analíticos. Segundo o Google Gemini Team (2023), o modelo demonstra proficiência nativa no processamento de caracteres e layouts variados, eliminando a necessidade de motores de OCR (*Optical Character Recognition*) auxiliares para a leitura de documentos digitalizados.

Em testes de desempenho padronizados, o Gemini atingiu resultados de estado da arte (*State of Art*) em *benchmarks* críticos para o setor corporativo, como o DocVQA

(*Visual Question Answering* em documentos) e o ChartQA (interpretação lógica de gráficos). Esses indicadores validam a robustez do modelo para tarefas que exigem raciocínio multimodal sobre dados visuais não estruturados (Google Gemini Team, 2023).

No ecossistema financeiro, essa tecnologia apresenta aplicação prática imediata na automação de processos de *Know Your Customer* (KYC) e análise de crédito. O modelo é capaz de extrair e validar dados de documentos oficiais (como RG, CNH e comprovantes de residência) e interpretar demonstrativos financeiros complexos (como gráficos de rentabilidade ou extratos em PDF), convertendo informações visuais em dados estruturados para a tomada de decisão bancária.

### 2.3.2.3 Compreensão Multimodal em Áudio (*Audio Understanding*)

A arquitetura de processamento sonoro do modelo Gemini fundamenta-se na integração do *Universal Speech Model* (USM) para a codificação de sinais de áudio em 16kHz. Segundo o Google Gemini Team (2023), essa abordagem permite que o sistema capture nuances paralinguísticas essenciais — como entonação, pausas e variações de tom — que transcendem a simples transcrição textual.

Diferentemente das arquiteturas convencionais que operam em "cascata" (sequenciando módulos de *Speech-to-Text*, LLM e *Text-to-Speech*), o Gemini processa o áudio como um *token* de entrada nativo na estrutura Transformers. Essa unificação elimina a perda de informações não verbais inerente aos processos de transcrição tradicionais, onde o contexto emocional é frequentemente descartado.

No contexto bancário, essa capacidade de *Audio Understanding* nativo é crítica para a análise de sentimento em tempo real. O modelo torna-se apto a identificar estados psicológicos do cliente, distinguindo, por exemplo, um tom de urgência ou frustração baseando-se na prosódia da fala. Tal funcionalidade permite modular a resposta do sistema de forma empática e contextualizada, priorizando atendimentos críticos sem intervenção humana imediata.

Além da compreensão, a arquitetura unificada otimiza a geração de resposta (*Text-to-Speech*), reduzindo a latência da interação e produzindo uma síntese vocal com ritmo e naturalidade superiores aos modelos sintéticos padrões. A eliminação das etapas intermediárias de conversão valida a viabilidade técnica da solução para um atendimento financeiro fluido e humanizado.

## 2.4 Prompt Engineering: A Ciência da Humanização e Controle

A eficácia de agentes de IA depende da precisão das instruções que regem seu comportamento, disciplina conhecida como *Prompt Engineering*. Definida pela Microsoft (2024) como a otimização de entradas para garantir aderência às regras

de negócio, essa prática permite adaptar *Large Language Models* (LLMs) a tarefas complexas sem a necessidade de re-treinamento (*fine-tuning*) (ZHANG et al., 2024).

A arquitetura do assistente “DUDA” fundamenta-se na combinação de quatro estratégias técnicas:

**System Persona Prompting:** Definição rígida do *System Role* como “Especialista em finanças do Bank AI”, encapsulando a identidade para assegurar consistência tonal e normativa.

**Context Grounding (RAG):** Imposição de prioridade absoluta aos dados recuperados do banco vetorial sobre o conhecimento paramétrico do modelo. Essa ancoragem inibe alucinações ao restringir as respostas à documentação fornecida.

**Hierarchical Reasoning:** Implementação de *Chain of Thought* (Cadeia de Pensamento) como uma árvore de decisão lógica. O modelo valida sequencialmente o contexto antes de responder, bloqueando inferências fora do domínio financeiro.

**Few-Shot Prompting:** Calibragem via injeção de exemplos práticos (*shots*) de “Entrada vs. Resposta Ideal”. A técnica padroniza o tratamento de casos de borda (*edge cases*) demonstrando os limites operacionais do agente.

Essa orquestração garante que o protótipo opere como um sistema determinístico e seguro, mitigando a imprevisibilidade estocástica inerente aos modelos generativos.

### 3 METODOLOGIA

A metodologia deste trabalho classifica-se como pesquisa aplicada de natureza exploratória. O estudo integra a revisão bibliográfica sobre o estado da arte em Inteligência Artificial Generativa com uma abordagem experimental, concretizada no desenvolvimento de um artefato de software. O objetivo central é validar a viabilidade técnica da substituição de menus de navegação estáticos por interfaces conversacionais fluídas no setor bancário.

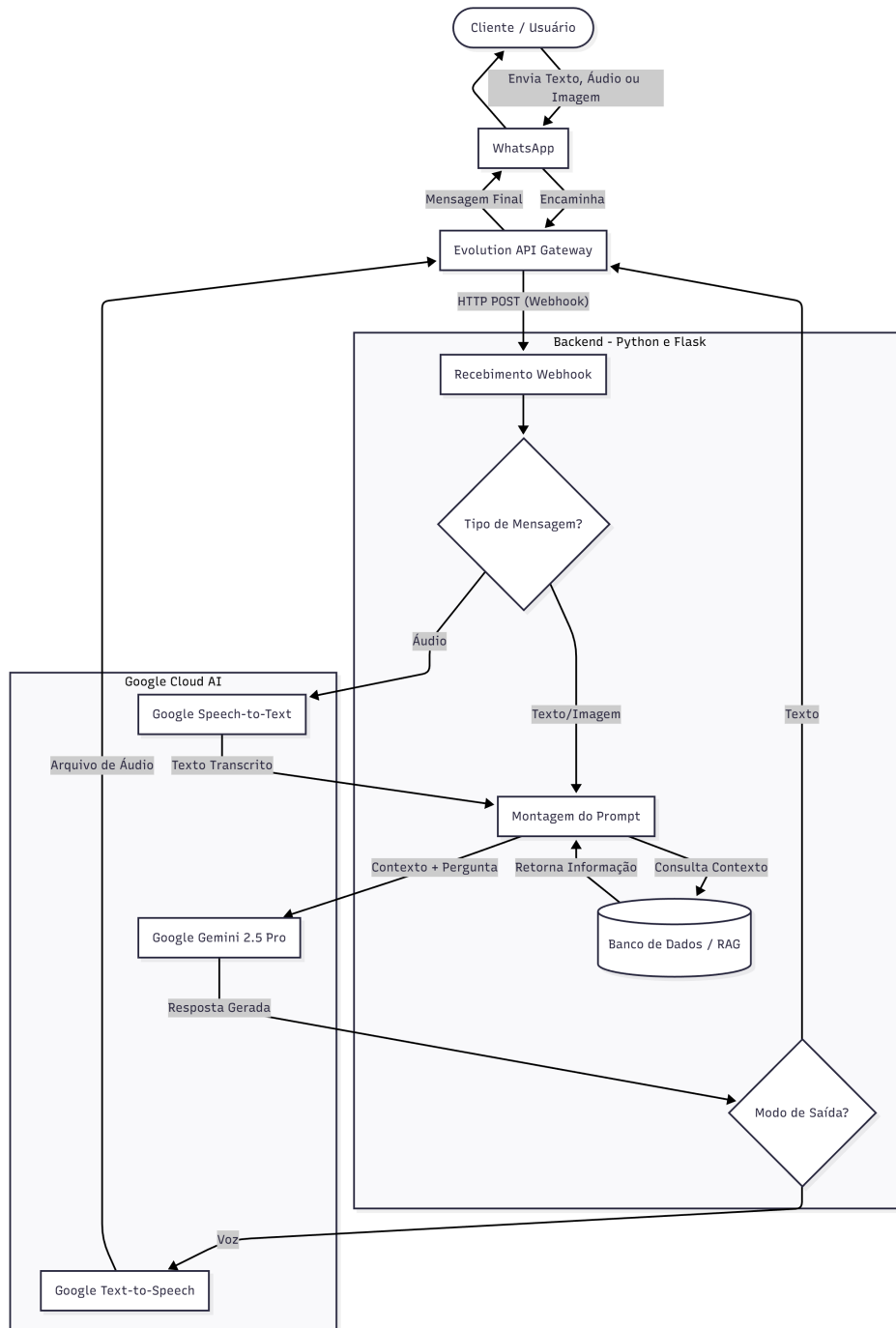
A seguir, detalha-se a arquitetura proposta, a pilha tecnológica e os fluxos de processamento de dados implementados.

#### 3.1 Arquitetura do Sistema

O protótipo foi desenvolvido sob uma arquitetura de microsserviços orientada a eventos. A comunicação entre a interface do usuário (WhatsApp) e o núcleo de processamento ocorre de forma assíncrona via *Webhooks*, intermediada pelo *gateway* Evolution API.

O fluxo de dados segue a estrutura apresentada na Figura 5, onde o servidor atua como orquestrador central das requisições junto às APIs do Google Cloud.

Figura 5 – Diagrama da arquitetura de fluxo de dados do sistema



Fonte: Elaborado pelo autor (2025).

### 3.2 Estrutura Tecnológica e Ferramentas

A implementação do sistema baseou-se nas seguintes escolhas técnicas<sup>1</sup>:

- **Linguagem de Programação:** Python, utilizado em conjunto com o *framework*

<sup>1</sup> O código-fonte completo e a documentação técnica estão disponíveis em: <<https://github.com/LordeVitor/Chatbot-com-Suporte-de-Voz-para-o-Setor-Banc-rio>>

Flask. A escolha justifica-se pela robustez no tratamento de requisições HTTP e pelo amplo ecossistema de bibliotecas voltadas para IA.

- **Persistência de Dados:** SQLite, operando em dupla função: armazenamento relacional do histórico de conversas e gerenciamento vetorial para o sistema RAG (*Retrieval Augmented Generation*).
- **Infraestrutura e Orquestração:** Docker e Docker Compose, utilizados para garantir o isolamento dos serviços, a portabilidade do ambiente e a consistência na execução entre diferentes máquinas.

### 3.3 Núcleo de Processamento e Pipelines de Dados

O núcleo do sistema implementa lógicas condicionais para tratamento de diferentes tipos de mídia. Para garantir a estabilidade, aplica-se inicialmente um **filtro de validação temporal**, descartando mensagens com latência superior a cinco minutos para prevenir o processamento de contexto obsoleto.

Após a validação, o processamento é segregado nos seguintes *pipelines*:

1. **Processamento de Texto:** As mensagens são higienizadas e enviadas à API Google Gemini. O *prompt* é enriquecido com o histórico da sessão e as definições de persona do sistema.
2. **Processamento de Áudio:** O sistema opera um fluxo *full-duplex* simulado. Arquivos de áudio (OGG) são transcritos via *Speech-to-Text* (STT). O texto resultante é processado pela IA e a resposta é sintetizada novamente em áudio via *Text-to-Speech* (TTS) antes do envio ao usuário.
3. **Processamento Multimodal:** Imagens e PDFs são analisados pela visão computacional do modelo Gemini. Implementou-se um *buffer* de estado: arquivos recebidos sem legenda são retidos temporariamente até que o usuário forneça uma instrução textual de contexto, garantindo a precisão da análise.

### 3.4 Gestão de Conhecimento e Controle

Para mitigar alucinações e assegurar a conformidade com normas bancárias, o sistema utiliza duas estratégias de controle:

- **RAG (Retrieval Augmented Generation):** Documentos técnicos são fragmentados e indexados vetorialmente. Antes da geração da resposta, o sistema recupera trechos semânticos relevantes e os injeta no contexto da IA.

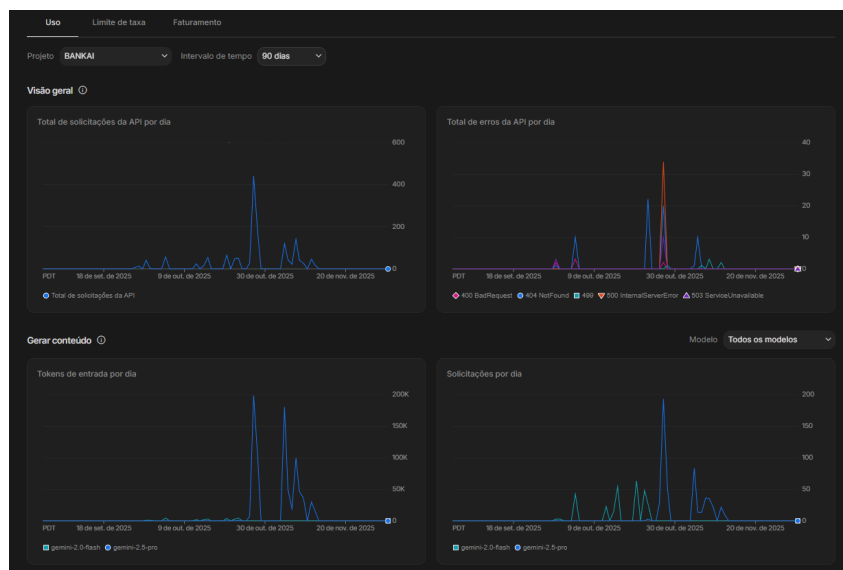
- **Controle Dinâmico:** Módulos administrativos permitem o envio de mensagens em massa (*broadcast*) e a alteração do comportamento do *bot* em tempo real (ex: alternância entre modo "Assistente" e "Vendas") sem necessidade de reinicialização do servidor.

### 3.5 Análise de Custos Operacionais

A infraestrutura financeira do projeto baseou-se no consumo de serviços da Google Cloud Platform (GCP). Os custos operacionais restringiram-se às requisições de APIs para transcrição de áudio (*Speech-to-Text*), síntese de voz (*Text-to-Speech*) e processamento de linguagem natural via modelo Gemini 2.5 Pro.

Durante o ciclo de desenvolvimento e testes (90 dias), o custo total acumulado para a operação do protótipo foi de R\$ 26,10, demonstrando baixo *overhead* financeiro. A discriminação dos valores segue a tabela oficial apresentada na Figura 6.

Figura 6 – Tabela de precificação do modelo Gemini 2.5 Pro



Fonte: Google AI Studio (2025).

## 4 DESENVOLVIMENTO E ANÁLISE EXPERIMENTAL

A etapa de desenvolvimento e validação expôs desafios técnicos inerentes à integração de sistemas heterogêneos. Esta seção detalha as limitações de infraestrutura encontradas e define o protocolo de testes utilizado.

### 4.1 Das Escolhas Técnicas e Suas Limitações

Para a interface de comunicação do protótipo, optou-se pela utilização da Evolution API (versão 1.8.2), selecionada por sua estabilidade documentada. No entanto, durante a fase de validação, identificou-se uma limitação técnica específica desta versão: uma anomalia na classificação de metadados de novos números de telefone.

O sistema apresentou falhas na distinção de identificadores, categorizando incorretamente novos usuários como LID (*Lightweight Identity*), uma regra de privacidade do protocolo WhatsApp. Essa inconsistência dificultou, em cenários específicos, o reconhecimento do remetente pelo algoritmo. Embora a correção exista na versão 2.3, a manutenção da versão 1.8.2 constituiu uma restrição técnica do ambiente de produção deste estudo de caso.

### 4.2 Protocolo de Testes e Massa de Dados

Para a validação funcional, estabeleceu-se um ambiente controlado alimentado por um conjunto de dados sintéticos, estruturados para simular operações bancárias reais e avaliar a eficácia dos módulos RAG e multimodal. O *dataset* de teste foi composto por três vetores principais:

1. **Base de Conhecimento (Contexto RAG):** Ingestão de um "Manual de Produtos" em formato textual (.txt), contendo regras de negócio restritivas:
  - *Lógica de Crédito:* Fórmula para cálculo de margem consignável (Vencimentos - Descontos Obrigatórios  $\times$  35%).
  - *Consultoria Financeira:* Diretrizes da regra "50-30-20" para categorização de orçamento.
  - *Portfólio:* Definições técnicas de produtos de investimento (Tesouro Selic, CDBs e FIs).
2. **Dados Não Estruturados (Multimodal):** Utilização de um arquivo PDF padronizado (Holerite\_Exemplo.pdf) para validar a extração de dados financeiros:
  - Vencimentos: R\$ 2.926,64.

- Descontos (INSS/IR): R\$ 232,89.
- Renda Líquida Calculada: R\$ 2.693,75.

### 3. Variáveis de Interação: Aplicação de *inputs* heterogêneos para teste de robustez:

- *Solicitações Informais*: Perguntas em linguagem natural (ex: "Como funciona o crédito?", "Categorize meus gastos").
- *Comandos de Ação*: Envios de arquivos e solicitações de cálculo.
- *Áudio em Cenários Variados*: Gravações com diferentes níveis de ruído para aferir a precisão da transcrição (STT).

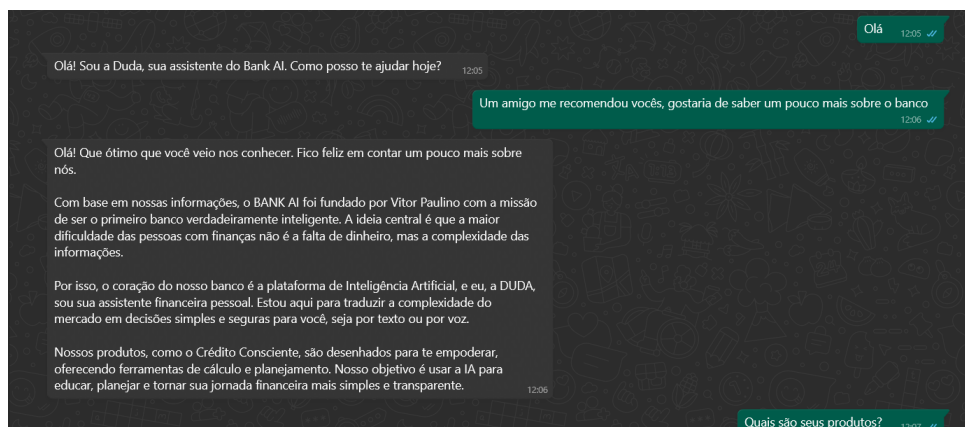
## 4.3 Estudo de Caso I: Interação Textual e Multimodal

Este cenário avalia a eficácia do agente em conduzir um ciclo completo de atendimento bancário via texto. O protocolo de teste verifica três pilares: a retenção de contexto conversacional; a aplicação de regras de negócio via RAG; e a capacidade de processamento de documentos não estruturados. A sequência completa das interações e evidências adicionais deste teste encontram-se documentadas externamente<sup>1</sup>.

### 4.3.1 Iniciação e Contextualização

O fluxo inicia-se com o estabelecimento da persona "DUDA", que identifica a intenção do usuário e apresenta o portfólio de serviços baseada na consulta à base de conhecimento (RAG). A Figura 7 demonstra a fluidez da linguagem natural na abordagem inicial, superando a rigidez de menus tradicionais.

Figura 7 – Interação inicial e definição de persona



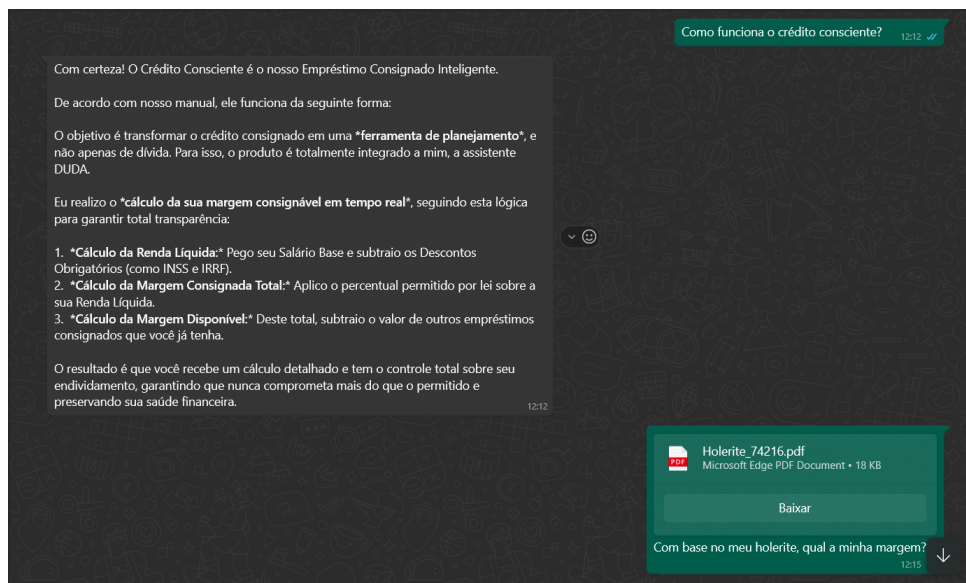
Fonte: Elaborado pelo autor (2025).

<sup>1</sup> O registro integral dos logs de conversação e capturas de tela suplementares (incluindo testes de instabilidade de rede e categorização de despesas) está disponível em: <[https://drive.google.com/drive/folders/10DSPWMuVJPzH\\_c5d6MutcDkVsBn7g310?usp=sharing](https://drive.google.com/drive/folders/10DSPWMuVJPzH_c5d6MutcDkVsBn7g310?usp=sharing)>

### 4.3.2 Processamento Multimodal e Lógica de Negócio

A capacidade crítica do sistema é evidenciada na etapa de análise de crédito. O usuário submete um comprovante de renda (holerite) em formato PDF sem metadados estruturados. O modelo Gemini realiza a extração dos campos financeiros (vencimentos e descontos) e aplica a fórmula de margem consignável definida nas regras de negócio, conforme visualizado na Figura 8.

Figura 8 – Processamento multimodal de PDF e cálculo de margem



Fonte: Elaborado pelo autor (2025).

Durante o ciclo de testes, validou-se também a robustez da persistência de estado. Simulações de instabilidade de rede (interrupções propositalmente de 5 minutos) demonstraram que o sistema é capaz de recuperar o histórico da sessão e retomar a explicação sobre "Gestão de Renda" e a aplicação da heurística "50-30-20" sem exigir a repetição de dados pelo usuário.

### 4.3.3 Consultoria Financeira e Fechamento

O atendimento evoluiu para uma etapa de consultoria personalizada. Com base nos dados extraídos do documento, o agente categorizou automaticamente as despesas do usuário e sugeriu metas financeiras. O ciclo encerra-se com a oferta contextualizada de um produto de investimento (Tesouro Selic) e a formalização da venda, validando a capacidade de *cross-selling* do sistema (Figura 9).



*Text* (STT). O texto resultante é processado pelo modelo Gemini, cuja resposta (gerada em Markdown) passa por um pré-processamento de SSML (*Speech Synthesis Markup Language*) no *backend* antes de ser sintetizada em áudio MP3 pela API *Text-to-Speech* (TTS).

Testes de campo indicaram que a qualidade da transcrição é sensível à relação sinal-ruído. Captações realizadas em ambientes com alta poluição sonora resultaram em perda de informação, apontando a necessidade futura de implementação de filtros de redução de ruído no pré-processamento.

#### **4.4.1 Análise de Transcrição e Diálogo**

A Tabela 1 apresenta a transcrição integral de uma interação realizada no ambiente de testes, evidenciando a capacidade do agente em processar cálculos financeiros complexos, interpretar documentos e conduzir o fluxo de vendas (contratação e KYC - *Know Your Customer*) inteiramente por voz.

Tabela 1 – Transcrição integral da interação de voz: Negociação de Crédito

| Interlocutor   | Conteúdo da Transcrição   |
|----------------|---|
| <b>Cliente</b> | Olá, tudo bem? Um amigo meu me recomendou vocês, ele falou que vocês têm a opção de pegar empréstimo consignado. Será que eu conseguiria fazer uma simulação com vocês?   |
| <b>Chatbot</b> | Claro! Para calcular sua margem consignável, preciso:<br>1 - Salário bruto;<br>2 - Descontos obrigatórios (INSS/IRRF);<br>3 - Parcelas de consignados atuais.   |
| <b>Cliente</b> | Esse é o meu holerite. Será que você poderia fazer o cálculo para mim?  |
| <b>Chatbot</b> | Claro, recebi seu holerite e já realizei a análise completa.<br>Renda líquida = R\$ 4.050 – R\$ 466,25 = R\$ 3.583,75.<br>Margem total (35%) = R\$ 1.254,31.<br>Menos parcela atual (R\$ 590):<br><b>Margem disponível = R\$ 664,31.</b><br>Quer ver as opções?           |
| <b>Cliente</b> | Gostaria sim. Pode me informar?   |
| <b>Chatbot</b> | Com certeza. Preparei algumas simulações para você ter uma ideia clara do valor com R\$ 664,31 de parcela:<br>24 meses: ~ R\$ 12.790<br>48 meses: ~ R\$ 21.340<br>72 meses: ~ R\$ 26.860<br>84 meses: ~ R\$ 28.000<br>Qual prefere?                                       |
| <b>Cliente</b> | Eu gostaria de simular mais ou menos um ano e meio.   |
| <b>Chatbot</b> | Claro. Vamos simular para o prazo de 18 meses (um ano e meio), Com R\$ 664,31 de parcela:<br><b>18 meses → valor liberado ~ R\$ 10.116.</b><br>Atende você?   |
| <b>Cliente</b> | Atendem sim. Acho que é um bom negócio.   |
| <b>Chatbot</b> | Excelente. Fico feliz em saber que a simulação atendeu às suas expectativas. O próximo passo é muito simples, vou:<br>1. Gerar contrato (R\$ 10.116 / 18x / R\$ 664,31).<br>2. Enviar para assinatura digital via email.<br>3. Após assinatura, crédito em até 24h úteis. |
| <b>Cliente</b> | Claro. Eu tenho que passar algum documento como um RG com foto?   |
| <b>Chatbot</b> | Ótima pergunta. Sim, a verificação de identidade é uma etapa fundamental e obrigatória. Você enviará:<br>Foto frente e verso do RG ou CNH;<br>Uma selfie para validação.  |
| <b>Cliente</b> | Ok, muito obrigado.   |
| <b>Chatbot</b> | Imagina, eu que agradeço pela confiança. Fico feliz em poder ajudar. O e-mail já foi enviado. Qualquer dúvida, me chame.  |

**Fonte:** Elaborado pelo Autor (2025).

#### 4.5 Gestão Administrativa e Controle Dinâmico

O sistema inclui um módulo de administração via CLI (*Command Line Interface*) para operações de *Broadcast* (envio em massa) e controle de estado da IA.

#### 4.5.1 Mensageria Ativa (Broadcast)

O painel permite o disparo de mensagens ativas em três modalidades: Geral, para toda a base; Específico, para usuários selecionados; e Personalizado, que utiliza injeção de variáveis (ex: {name}) para customizar o conteúdo. A Figura 11 exibe a interface de seleção para envio personalizado.

Figura 11 – Interface de envio de mensagens personalizadas

```

--- Modo Atual da IA: SALES ---

--- Painel de Controlo ---
1. Enviar a mesma mensagem para TODOS
2. Enviar mensagem personalizada para TODOS
3. Enviar mensagem para contactos ESPECÍFICOS
4. Alterar Modo da IA (Vendas/Padrão)
5. Sair
Escolha uma opção: 2

Escreva a sua mensagem. Use '{name}' onde quiser que o nome do utilizador apareça.
Template da mensagem: Olá {name}, este é um teste do protótipo DUDA

```

Fonte: Elaborado pelo autor (2025).

#### 4.5.2 Alteração Dinâmica de Persona

Através da técnica de *Prompt Engineering*, o administrador pode alternar o comportamento do agente em tempo real, sem reinicialização do servidor. A funcionalidade permite a transição entre o "Modo Padrão"(Generalista) e o "Modo Vendas"(Especialista em Finanças), conforme exibido na Figura 12.

Figura 12 – Painel de alteração do modo de operação (Standard vs Sales)

```

PS C:\Users\Vitor\Desktop\tcc backup\Tcc prog> py sender.py

--- Modo Atual da IA: STANDARD ---

--- Painel de Controlo ---
1. Enviar a mesma mensagem para TODOS
2. Enviar mensagem personalizada para TODOS
3. Enviar mensagem para contactos ESPECÍFICOS
4. Alterar Modo da IA (Vendas/Padrão)
5. Sair
Escolha uma opção: 4

--- Alterar Modo da IA ---
1. Modo Vendas (Foco em produtos financeiros)
2. Modo Padrão (Assistente geral)
Escolha o novo modo:

```

Fonte: Elaborado pelo autor (2025).

A distinção comportamental entre os modos é regida por diretrizes rígidas de sistema, conforme comparativo abaixo:

- **Persona Financeira (RAG):** Prioriza o contexto recuperado da base de conhecimento. Possui diretrizes de *fallback* para recusar perguntas fora do tópico financeiro e focar estritamente em cálculos e produtos bancários.
- **Persona Standard (Generalista):** Opera com liberdade criativa, habilitada para tarefas diversas (resumos, análises gerais) e interações casuais, evitando recusas desnecessárias.

#### 4.6 Aplicação do Retrieval-Augmented Generation (RAG)

Para mitigar alucinações e assegurar a precisão factual das respostas financeiras, a arquitetura do sistema incorpora o paradigma de *Retrieval-Augmented Generation* (RAG). Essa implementação permite que o agente consulte dinamicamente uma base de conhecimento proprietária antes de gerar qualquer resposta, garantindo que as orientações (como cálculos de margem ou regras de investimento) estejam em estrita conformidade com as políticas do banco fictício.

O fluxo de processamento do RAG no protótipo ocorre em quatro etapas sequenciais:

**Ingestão e Indexação:** Documentos regulatórios e manuais de produtos (em formato PDF e texto) foram fragmentados e convertidos em vetores (*embeddings*), armazenados no banco de dados para busca eficiente.

**Recuperação Semântica (*Retrieval*):** Ao receber uma consulta do usuário (seja via texto direto ou transcrição de áudio), o sistema converte a pergunta em um vetor e realiza uma busca por similaridade na base indexada, recuperando os trechos mais relevantes do manual.

**Enriquecimento de Contexto (*Augmentation*):** Os trechos recuperados são injetados no *prompt* do sistema, instruindo o modelo Gemini a utilizar exclusivamente aquelas informações como "fonte da verdade".

**Geração Assistida:** O modelo processa o *prompt* enriquecido e gera a resposta final, garantindo que o cálculo ou a informação fornecida esteja fundamentada nos dados internos recuperados, e não apenas no conhecimento pré-treinado da LLM.

Esse mecanismo é representado visualmente no fluxograma geral do sistema, detalhado na seção anterior, onde o módulo de banco de dados interage diretamente com o núcleo de processamento antes da chamada à API do Gemini.

## 5 DISCUSSÕES DOS RESULTADOS

A análise dos dados obtidos nas simulações práticas valida a hipótese central deste trabalho: a aplicação de Grandes Modelos de Linguagem (LLMs) em arquiteturas bancárias supera as limitações dos sistemas determinísticos baseados em regras. Os resultados indicam uma mudança de paradigma na interface homem-máquina, movendo-se de navegações estáticas para interações fluidas e contextuais.

### 5.1 Análise da Interação Textual e Semântica

O estudo de caso textual demonstrou que a arquitetura proposta sustenta três pilares fundamentais para a automação bancária moderna:

- **Persistência de Estado e Coerência:** O sistema exibiu robustez na manutenção do contexto conversacional (*Context Window*). Mesmo diante de interrupções de rede simuladas, o agente recuperou o estado da transação sem exigir redundância de dados por parte do usuário. Essa característica elimina a fricção comum em *chatbots* legados, onde falhas de conexão reiniciam o fluxo de atendimento.
- **NLU e Extração de Entidades:** A capacidade de *Natural Language Understanding* (NLU) provou-se eficaz na identificação de intenções complexas (ex: "simulação de consignado") e na extração proativa de entidades (salário base, descontos). O modelo não apenas interpretou a sintaxe, mas aplicou a lógica de negócio definida no manual de conhecimento para solicitar dados faltantes, mimetizando a conduta de um analista humano.
- **Consultoria Dinâmica:** Diferente de sistemas que apenas recuperam informações estáticas, o protótipo atuou como um consultor financeiro. A capacidade de gerar cenários de crédito baseados nas variáveis individuais de cada cliente (personalização em tempo real) representa um diferencial competitivo, transformando o suporte operacional em uma ferramenta de retenção e vendas.

### 5.2 Análise da Interação Vocal e Multimodalidade

A avaliação do fluxo de voz evidenciou que a orquestração sequencial (STT → LLM → TTS) é capaz de entregar latências aceitáveis para uma experiência de conversação quase natural. A modularidade do sistema permitiu a alternância fluida entre canais: quando o fluxo vocal exigiu a análise visual de um documento (PDF), o agente processou a entrada multimodal sem quebrar a continuidade do diálogo.

Um achado relevante foi a resiliência do sistema em cenários de ruído acústico. Ao identificar falhas na transcrição (STT) devido a interferências externas, o agente acionou mecanismos de recuperação de erro (*fallback*), solicitando a repetição da instrução de forma polida. Esse comportamento reforça a antropomorfização da interface, aproximando a interação da máquina à de um atendente humano.

### 5.3 Viabilidade Técnica e Operacional

Os experimentos confirmam a viabilidade da implementação, destacando-se quatro implicações estratégicas para o setor financeiro:

- **Eficiência Operacional:** A automação de cálculos complexos (Margem Consignável) e a extração automática de dados via OCR reduzem drasticamente o Tempo Médio de Atendimento (TMA) e mitigam riscos de erro humano na entrada de dados.
- **Conformidade e Segurança:** A inclusão de etapas de verificação de identidade (KYC) e a explicabilidade das solicitações alinham o protótipo aos requisitos de transparência preconizados pela Lei Geral de Proteção de Dados (LGPD) e pelo Marco Legal da IA (PL 2.338/2023).
- **Alta Disponibilidade (24/7):** A arquitetura baseada em nuvem supera a limitação de horário comercial das agências físicas, oferecendo atendimento primário resolutivo a qualquer momento, atendendo à demanda por onipresença dos serviços bancários digitais.
- **Humanização da Interface:** A supressão da "robotização", alcançada pela fluidez dos modelos generativos e pela síntese de voz natural, eleva o índice de satisfação do usuário (*CSAT*), posicionando a solução não apenas como uma ferramenta funcional, mas como um ativo de relacionamento.

## 6 TRABALHOS FUTUROS

A validação experimental do protótipo confirmou a viabilidade técnica da aplicação, mas também delineou o roteiro necessário para a evolução do artefato acadêmico em direção a um produto de mercado. As frentes de investigação sugeridas visam mitigar as latências observadas e validar a sustentabilidade econômica do modelo.

### 6.1 Evolução da Arquitetura de Software

- **Telefonia Síncrona (VoIP):** Substituição do modelo assíncrono (gravar/enviar) por canais de voz em tempo real via APIs especializadas. Isso reduzirá drasticamente a latência, permitindo interrupções naturais na fala (*barge-in*) e aproximando a interação da dinâmica humana.
- **Robustez Acústica (DSP):** Implementação de algoritmos de Processamento Digital de Sinais no pré-processamento. A aplicação de filtros de cancelamento de ruído e normalização de áudio antes da transcrição (STT) é crítica para garantir a eficácia do sistema em ambientes não controlados.
- **Integração com CRM:** Evolução do assistente para uma ferramenta proativa de vendas. A conexão via API com sistemas de CRM permitirá o acesso ao histórico transacional em tempo real, habilitando a qualificação automática de *leads* e a personalização de ofertas baseada em risco de crédito.
- **Otimização de Performance:** Adoção de *Audio Streaming* para reduzir o tempo de percepção da resposta (*Time to First Byte*) e implementação de *Cache Semântico*. Esta estratégia evita o reprocessamento de dúvidas frequentes, otimizando o consumo de tokens e reduzindo custos operacionais.

### 6.2 Validação Estratégica e de Negócio

- **Análise Econômica (TCO e ROI):** Condução de estudo comparativo entre o Custo Total de Propriedade da infraestrutura de IA e os custos de *Call Centers* tradicionais. A modelagem deve projetar o retorno sobre o investimento considerando diferentes precificações de API para validar a escalabilidade financeira.
- **Métricas de Usabilidade:** Execução de testes de campo com usuários reais para mensurar indicadores qualitativos (*NPS, CSAT*). O objetivo é comparar estatisticamente a eficácia resolutiva do agente de IA frente a operadores humanos, validando a aceitação social da tecnologia.

## 7 TRABALHOS RELACIONADOS

A validação da arquitetura proposta neste estudo encontra respaldo em implementações consolidadas no mercado financeiro. A análise de assistentes virtuais em larga escala demonstra que a demanda por interações automatizadas e multimodais é um requisito operacional estabelecido globalmente.

### 7.1 Agentes Virtuais em Escala: Erica (BofA) e BIA (Bradesco)

No cenário internacional, o assistente Erica, do Bank of America, constitui o *benchmark* do setor, ultrapassando 2 bilhões de interações em 2024 (Bank of America, 2024). Sob a ótica da engenharia de software, o caso Erica valida dois pilares centrais do nosso protótipo: a multimodalidade nativa, capaz de processar inputs heterogêneos (voz/texto), e a hiperpersonalização, análoga à nossa funcionalidade de consultoria financeira baseada em regras.

No Brasil, a BIA (Bradesco Inteligência Artificial) representa o estado da arte com mais de 2,2 bilhões de interações (Banco Bradesco S.A., 2023). Um diferencial técnico relevante da BIA é sua atuação híbrida, servindo também como suporte à decisão para gerentes ao consultar regimentos internos. Esta característica corrobora a escolha da arquitetura RAG (*Retrieval Augmented Generation*) utilizada neste trabalho: assim como a BIA consulta normas corporativas, nosso protótipo acessa o "Manual de Produtos" para garantir precisão técnica e mitigar alucinações.

### 7.2 Lacuna de Mercado e Aplicabilidade Proposta

Enquanto Erica e BIA constituem soluções proprietárias de alto custo desenvolvidas por grandes conglomerados, este trabalho identifica uma lacuna estratégica no setor de crédito terceirizado (correspondentes bancários e telemarketing). Este nicho, caracterizado por alta rotatividade de pessoal e margens operacionais comprimidas, historicamente carece de acesso a tecnologias de ponta.

O protótipo desenvolvido propõe a democratização dessas capacidades através do uso de LLMs comerciais (Gemini) e orquestração de baixo custo. A solução preenche o vácuo tecnológico existente entre o atendimento humano precário e as IAs corporativas de elite, oferecendo a financeiras de médio porte a mesma sofisticação técnica — disponibilidade 24/7, personalização em massa e precisão normativa — necessária para competir na economia digital.

## 8 Considerações Finais

O presente trabalho investigou a viabilidade da aplicação de tecnologias de Inteligência Artificial, com ênfase em recursos de *text-to-speech* e *speech-to-text*, na transformação do atendimento bancário em uma experiência mais natural, personalizada e acessível. Por meio de uma abordagem que combinou pesquisa bibliográfica exploratória e o desenvolvimento de um protótipo funcional, foi possível analisar não apenas os aspectos teóricos e regulatórios da IA no setor financeiro, mas também validar empiricamente a eficácia dessas tecnologias em cenários práticos de atendimento.

Os resultados obtidos demonstraram que a integração de *Large Language Models* (LLMs) multimodais, como o Google Gemini, com tecnologias de conversão de voz (STT/TTS), viabiliza a criação de sistemas de atendimento digital capazes de superar as limitações dos chatbots tradicionais baseados em regras. O protótipo desenvolvido apresentou capacidade robusta de manutenção de contexto conversacional, compreensão de linguagem natural, processamento multimodal de documentos e personalização de respostas, características essenciais para atender às demandas do setor bancário moderno.

Do ponto de vista regulatório, o trabalho evidenciou a importância do alinhamento das soluções de IA com as diretrizes propostas pelo Marco Regulatório da Inteligência Artificial (PL 2.338/2023) e pela Lei Geral de Proteção de Dados (LGPD). A transparência no caráter automatizado das interações, a garantia de revisão humana em decisões críticas, a proteção de dados pessoais e a prevenção de vieses discriminatórios são pilares fundamentais para a adoção responsável e sustentável dessas tecnologias no setor financeiro.

Os Estudos de Caso práticos realizados com o protótipo evidenciaram tanto as potencialidades quanto às limitações das tecnologias empregadas. A interação textual demonstrou fluidez e capacidade de personalização, enquanto a interação vocal revelou um nível surpreendente de naturalidade, ainda que com desafios relacionados à interpretatividade emocional do texto sintetizado. A capacidade multimodal de processar documentos em PDF e imagens, aliada à manutenção do contexto conversacional mesmo diante de instabilidades de rede, reforça a robustez da solução proposta.

Em síntese, este trabalho contribui para a compreensão de como agentes de IA podem redefinir a interação entre instituições financeiras e clientes, criando experiências mais humanas, inclusivas e eficientes no contexto digital. A viabilidade técnica, regulatória e mercadológica demonstrada pelo protótipo indica que o futuro do atendimento bancário está intrinsecamente ligado à adoção de tecnologias de IA multimodais, capazes de oferecer disponibilidade ininterrupta (*anytime*), personalização em escala e, sobretudo, uma experiência verdadeiramente humanizada.

## Referências

- Banco Bradesco S.A. *IA Confiável (Trusted AI) - Compromisso do Bradesco com a IA confiável*. São Paulo, 2023. Disponível em: <<https://assets.bradesco/content/dam/portal-bradesco/assets/classic/pdf/sustentabilidade/trusted-ai.pdf>>.
- Bank of America. *BofA's Erica surpasses 2 billion interactions, helping 42 million clients since launch*. 2024. Press release. Disponível em: <<https://newsroom.bankofamerica.com/content/newsroom/press-releases/2024/04/bofa-s-erica-surpasses-2-billion-interactions-helping-42-million.html>>.
- FUSCALDO, D. *Bank of America's virtual assistant now has more than 10 million users*. Forbes, 2019. Disponível em: <<https://www.forbes.com/sites/donnafusco/2019/12/11/bank-of-americas-virtual-assistant-now-has-more-than-10-million-users/>>.
- Google Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. Disponível em: <<https://arxiv.org/abs/2312.11805>>.
- LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In: CURRAN ASSOCIATES, INC. *Conference on Neural Information Processing Systems (NeurIPS)*. 2020. v. 33, p. 9459–9474. Disponível em: <<https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>>.
- Microsoft. *Prompt Engineering Concepts in Azure AI Foundry*. 2024. Disponível em: <<https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/prompt-engineering>>.
- Reuters. *BofA says its virtual financial assistant used more than 2 billion times*. 2024. Disponível em: <<https://www.reuters.com/business/finance/bofa-says-its-virtual-financial-assistant-used-more-than-2-billion-times-2024-04-08/>>.
- RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 4. ed. Hoboken: Pearson, 2021.
- VARSHNEY, T. *Introduction to LLM Agents*. 2023. Acessado em: 22 dez. 2024. Disponível em: <<https://developer.nvidia.com/blog/introduction-to-llm-agents/>>.
- VASWANI, A. et al. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. Disponível em: <<https://arxiv.org/abs/1706.03762>>.
- ZHANG, Y. et al. A survey on prompt engineering for large language models: Methods and applications. *arXiv preprint arXiv:2412.05127*, 2024.