JNT FACIT BUSINESS AND TECHNOLOGY JOURNAL
ISSN: 2526-4281 - ANO 2026 - MÊS DE ABRIL
Ed. 73. Vol. 1 - Págs. 307-322
DOI: 10.5281/zenodo.19570699

CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

QUALIS A2

SISTEMA DE RECOMENDAÇÃO FINANCEIRA BASEADO EM IA PARA CONTROLE DE GASTOS[1]

AI-BASED FINANCIAL RECOMMENDATION SYSTEM FOR EXPENSE MITIGATION

João Victor Póvoa FRANÇA
Universidade Estadual do Tocantins (UNITINS)
E-mail: joaopovoa@unitins.br/joaovictorpfr@gmail.com
ORCID: http://orcid.org/0009-0007-6441-8827

Alex COELHO
Universidade Estadual do Tocantins (UNITINS)
E-mail: alex.c@unitins.br/alex.coelho@mail.uft.edu.br
ORCID: http://orcid.org/0000-0002-1576-7242

RESUMO

No contexto da gestão financeira pessoal, o controle de despesas enfrenta o desafio do atrito cognitivo e do excesso de informações, demandando soluções tecnológicas preditivas. O objetivo deste trabalho é apresentar a implementação de um Sistema de Recomendação Financeira Pessoal Inteligente, baseado em agentes de Inteligência Artificial, focado na mitigação de gastos desnecessários. A metodologia consistiu em uma pesquisa aplicada com desenvolvimento de software, estruturando uma arquitetura híbrida. Aplicaram-se heurísticas matemáticas determinísticas, como similaridade de *strings* (Distância de Levenshtein) e cálculo de médias móveis, para a mineração autônoma da "inércia de gastos". Em seguida, orquestrou-se um Modelo de Linguagem de Grande Escala via injeção de contexto para atuar como motor comportamental baseado na Teoria do *Nudge*. Os resultados evidenciaram o êxito operacional da solução: o isolamento do processamento reduziu custos computacionais e anulou o risco de "alucinações" matemáticas pela IA. O sistema validou sua capacidade de emitir conselhos explicáveis e estruturados, cruzando anomalias de consumo com metas financeiras em tempo real. Conclui-se que a transição de um paradigma descritivo, baseado em painéis estáticos, para um ecossistema de recomendação ativa e preditiva diminui drasticamente a carga cognitiva do usuário. A ferramenta consolida-se como um apoio robusto à decisão,

[1] COMO CITAR: (ABNT): FRANÇA, J. V. P.; COELHO, A. Sistema de Recomendação Financeira Baseado em IA Para Controle de Gastos. **JNT Facit Business and Technology Journal.** Qualis A2. ISSN: 2526-4281, Mês de Abril de 2026 - Ed. 73. VOL. 01. Págs. 307-322. Disponível: http://revistas.faculdadefacit.edu.br. Acesso em: ___/___/___.

adaptando-se às rotinas individuais e superando as limitações das aplicações de mercado no desafio da educação financeira contínua.

Palavras-chave: Finanças pessoais. Agentes de Inteligência Artificial. Sistemas de recomendação. Teoria do Nudge. Arquitetura de software.

ABSTRACT

In the context of personal financial management, expense control faces the challenge of cognitive friction and information overload, demanding predictive technological solutions. The objective of this study is to present the implementation of an Intelligent Personal Financial Recommendation System, based on Artificial Intelligence agents, focused on mitigating unnecessary expenses. The methodology consisted of applied research with software development, structuring a hybrid architecture. Deterministic mathematical heuristics, such as string similarity (Levenshtein Distance) and moving averages calculations, were applied for the autonomous mining of "spending inertia". Subsequently, a Large Language Model was orchestrated via context injection to act as a behavioral engine based on Nudge Theory. The results demonstrated the operational success of the solution: the isolation of processing reduced computational costs and eliminated the risk of mathematical "hallucinations" by the AI. The system validated its ability to issue explainable and structured advice, cross-referencing consumption anomalies with financial goals in real-time. It is concluded that the transition from a descriptive paradigm, based on static dashboards, to an active and prescriptive recommendation ecosystem drastically decreases the user's cognitive load. The tool consolidates itself as a robust decision-support system, adapting to individual routines and overcoming the limitations of market applications in the challenge of continuous financial education.

Keywords: Personal finance. Artificial Intelligence agents. Recommendation systems. Nudge theory. Software architecture.

INTRODUÇÃO

O controle e a gestão das finanças pessoais representam desafios contínuos para indivíduos que buscam aumentar a sua estabilidade econômica, garantir a segurança patrimonial e alcançar metas de curto e longo prazo. No contexto econômico contemporâneo, a supervisão constante dos próprios recursos e a

contenção de gastos desnecessários são essenciais para evitar o endividamento e melhorar a qualidade de vida.

Entretanto, avaliações internacionais recentes demonstram que grande parte da população enfrenta sérias dificuldades de organização financeira. Um relatório da OCDE (Organização para a Cooperação e Desenvolvimento Econômico) evidenciou que o Brasil possui uma das piores médias de educação financeira do mundo, refletindo uma baixa literacia que compromete a tomada de decisão e aumenta a exposição a comportamentos de risco (OECD, 2024). Corroborando esse cenário, o Banco Central do Brasil destaca que o fácil acesso ao crédito, aliado à falta de disciplina, impulsiona o consumismo e as compras por impulso (Banco Central Do Brasil, 2024).

Uma razão estrutural para essa dificuldade reside na psicologia do consumo. O excesso de opções e o bombardeio de ofertas no ambiente digital aumentam o atrito cognitivo, tornando mais difícil priorizar o essencial. Em um estudo clássico, Iyengar e Lepper (2000) documentaram que a sobrecarga de escolhas pode reduzir a ação e a satisfação do indivíduo, levando a decisões financeiras irracionais.

Neste aspecto, as tecnologias passaram a contribuir de maneira vital para as organizações e pessoas físicas, uma vez que possibilitam a automação de processos de triagem e análise de dados. Diversas empresas oferecem soluções para controle de despesas, mas que, entretanto, limitam-se a exibir dashboards estáticos e categorizações manuais, falhando em antecipar riscos ou fornecer orientações proativas.

Com o amadurecimento da Inteligência Artificial (IA) e do *Machine Learning*, novas perspectivas podem ser exploradas. A utilização de sistemas de recomendação baseados em IA permite que o comportamento do usuário seja mapeado para além da simples visualização de dados (Takahashi; Junior, 2015). O uso de modelos matemáticos de previsibilidade possibilita o cálculo de tendências a partir da análise de períodos de tempo, extraindo regras de recorrência, réplicas de comportamento e rotinas de consumo.

Segundo Zhang e Chen (2020), em domínios sensíveis como o financeiro, os sistemas de recomendação não devem apenas sugerir ações, mas precisam ser "explicáveis". Ou seja, para que o usuário confie na ferramenta e mude seus hábitos, a IA deve justificar matematicamente o porquê de uma recomendação. Como exemplo: "Você tem uma recorrência de gastos com delivery aos finais de semana que, neste período de 30 dias, compromete 15% da sua renda".

Assim, o presente trabalho detalha a implementação e validação prática de um sistema de recomendação financeira pessoal com o uso de IA. O foco central é a construção de insights financeiros automatizados, sustentados por um modelo de previsibilidade matemática que estuda os costumes e a rotina do usuário juntamente com a inteligência artificial.

METODOLOGIA

O desenvolvimento e validação da solução proposta no trabalho, consistiu em uma pesquisa aplicada. Segundo Silva e Menezes (2005), esse tipo de pesquisa busca gerar conhecimentos para resolver problemas práticos e específicos. Isso significa que não apenas estuda-se o problema do descontrole financeiro, mas cria-se uma ferramenta real e em formato de *software* para combatê-lo.

Mapeamento Sistemático da Literatura - MSL

Para entender como a tecnologia já vem sendo usada para ajudar nas finanças pessoais, foi realizado um mapeamento de trabalhos acadêmicos. O objetivo dessa busca foi, além de encontrar soluções semelhantes na área da tecnologia, e também encontrar oportunidades e limitações nas soluções já existentes, usando essas informações para construir uma proposta que cubra tais lacunas.

Ao encontrar os trabalhos relacionados, foram definidos critérios de busca com termos em inglês para garantir um alcance global, propondo-se como string de busca: ("personal finance" *OR* "expense management") *AND* ("recommender systems" *OR* "artificial intelligence"). As buscas foram realizadas em bases de dados reconhecidas na área de tecnologia, como Google Acadêmico, Scopus e IEEE Xplore.

A fim de assegurar que o levantamento refletisse apenas as tecnologias contemporâneas de Inteligência Artificial, aplicou-se um rigoroso recorte temporal, restringindo a análise a trabalhos publicados de 2019 em diante. O processo seguiu um funil metodológico em quatro etapas:

1. **Identificação Inicial (Aprox. 840 resultados):** Retorno bruto das bases de dados a partir da *string* de busca e do filtro temporal (2019-2026);
2. **Triagem de Relevância (Aprox. 315 trabalhos):** Descartando sistemas focados em finanças corporativas, criptomoedas ou que não abordem o gerenciamento de despesas do usuário;
3. **Filtro de Elegibilidade Rigoroso (12 trabalhos):** Trabalhos relacionados, e com mínima relevância a partir do tema financeiro;

4. **Aplicabilidade da IA (4 trabalhos)**. Descartaram-se os modelos de "caixa-preta" (*black-box*), mantendo apenas sistemas capazes de justificar as suas recomendações financeiras inteligentes;
5. **Seleção Comparativa (2 trabalhos):** Isolamento dos estudos mais aderentes à proposta para compor a análise comparativa direta do diferencial arquitetural.

Análise dos Trabalhos Relacionados

A partir da filtragem rigorosa do MSL, dois trabalhos recentes foram isolados para uma comparação direta, por representarem o atual estado da arte na intersecção entre Inteligência Artificial, gestão financeira e comportamento do usuário.

O primeiro trabalho selecionado é o de Wang *et al.* (2023), que propõe um sistema de categorização de despesas pessoais utilizando *Large Language Models* (LLMs).

- **Valências:** A principal força deste estudo reside na alta precisão do processamento de linguagem natural para classificar transações bancárias com descrições ambíguas.
- **Limitações (Gaps):** A arquitetura proposta submete todo o volume de dados brutos (milhares de linhas de extrato) diretamente à API do modelo inteligente. Isso resulta em um alto custo computacional (*tokens*) e elevada latência. Além disso, o sistema possui caráter estritamente descritivo; no qual informa ao usuário onde o dinheiro foi gasto, mas emite notificações passivas que exigem interpretação analítica por parte do indivíduo, falhando em estimular a mudança de hábito.

O segundo trabalho de destaque é o de **Ferreira e Santos (2024)**, que investigam a aplicação de gatilhos comportamentais (Teoria do *Nudge*) em aplicativos financeiros *mobile* para reduzir compras por impulso.

- **Valências:** O estudo é extremamente bem-sucedido ao comprovar que alertas visuais e arquitetura de escolhas reduzem o atrito cognitivo e engajam o usuário a economizar.
- **Limitações (Gaps):** A implementação tecnológica depende inteiramente de regras estáticas e programadas (*hardcoded*). O sistema dispara alertas baseados em limites fixos e pré-estabelecidos pelo usuário, carecendo de adaptação dinâmica. Quando o contexto do usuário muda, os alertas tornam-

se repetitivos e perdem a eficácia (fadiga de notificação), pois o sistema não consegue gerar explicações contextualizadas para os desvios.

Mapeamento de Mercado e a Teoria do *Nudge*

A exigência de categorização manual e essa postura puramente passiva geram alto atrito cognitivo, fator que frequentemente leva ao abandono dessas ferramentas por parte dos usuários (Febraban, 2023). A principal lacuna explorada no desenho do trabalho aqui proposto foi exatamente a transição desse modelo de "visualização passiva" para um modelo de "recomendação ativa e preditiva".

Essa mudança de paradigma encontra forte embasamento na Teoria do *Nudge* (Arquitetura de Escolhas), popularizada por Thaler e Sunstein (2021). Segundo os autores, as decisões melhoram quando o ambiente fornece estímulos simplificados, substituindo exigências pesadas por intervenções sutis. Para ilustrar essa diferença conceitual e prática, o Quadro 2 detalha o contraste entre os métodos tradicionais de mudança de comportamento e a abordagem facilitadora do *Nudge*.

Quadro 2: Comparativo entre Métodos de Mudança Impositiva e Técnicas de *Nudge*.

Aspecto de Análise	Mudança Impositiva (Tradicional)	Técnicas de Nudge (Arquitetura de Escolhas)
Natureza da Abordagem	Direta, óbvia e drástica. Baseada na imposição de regras, leis ou restrições de comportamento.	Indireta, sutil e tática. Baseada na facilitação e na reorganização do ambiente de escolha.
Mecanismo de Ação	Instrução explícita, aplicação de penalidades, fiscalização e exigência de alto esforço consciente.	Encorajamento, simplificação do ambiente, uso de exemplos e alinhamento com normas sociais.
Impacto Psicológico e Reação	Frequentemente confrontacional e ameaçadora; gera alto atrito cognitivo e propensão à resistência.	Cooperativa, fluída e menos ameaçadora; as opções tornam-se fáceis de imaginar e de executar.
Foco de Intervenção	Tentativa de forçar o indivíduo a mudar seu comportamento diretamente, ignorando tendências instintivas.	Modificação do ambiente (contexto) para que o indivíduo escolha a mudança de forma voluntária.

Exemplo: Finanças	Exigir o preenchimento manual de planilhas complexas de orçamento semanal.	Cesta pequena de compras no mercado para limitar gastos por impulso físicos.

Fonte: Adaptado de Thaler e Sunstein (2008).

Deste modo, a solução proposta no trabalho não apenas prevê o comportamento futuro com base nos hábitos matemáticos de consumo, mas também entrega uma ação clara e justificada. Em vez de uma notificação passiva, a ferramenta atua como um conselheiro proativo (exemplo: "Se você reduzir os pedidos de *delivery* no fim de semana em 15%, alcançará sua meta de poupança no prazo estipulado").

Contudo, para que essa transição de um simples "diário de gastos" para um motor preditivo fosse possível, a construção do *software* exigiu um rigoroso planejamento arquitetural. Foi necessário transcender o desenvolvimento tradicional de aplicativos e adotar um ecossistema focado na mineração de dados e na integração orquestrada com modelos de Inteligência Artificial.

Técnicas e Ferramentas Utilizadas

Para o atendimento do objetivo do trabalho, adotou-se uma arquitetura baseada em microsserviços e processamento em múltiplas camadas, priorizando a integridade dos dados, a estruturação matemática e a integração orquestrada com modelos de linguagem.

Essa padronização técnica é fundamental, pois dados bancários crus frequentemente apresentam formatos inconsistentes, e a injeção de dados ruidosos compromete a precisão dos algoritmos preditivos, induzindo a Inteligência Artificial a gerar respostas imprecisas ou "alucinações". As seções seguintes mostram as etapas do processo técnico utilizado.

- **Interface dos dados:** Utilizou-se a tipagem estática do TypeScript, para estabelecer a padronização absoluta das entradas. Essa estruturação técnica é fundamental para assegurar que o sistema receba apenas cargas de dados no formato esperado, blindando a aplicação contra inconsistências e eliminando ruídos antes que as informações alcancem a camada de processamento da Inteligência Artificial.
- **Regras de negócio**: Na camada de regra de negócio, a engenharia do sistema isolou o processamento computacional em um *service* de domínio específico, projetado para atuar como o Agente Inteligente da arquitetura, este service isola o processamento matemático, aplicando as heurísticas determinísticas

de forma preliminar, atuando de forma totalmente independente e anterior à chamada do LLM.

- **Mineração de dados**: Utiliza-se algoritmos de similaridade de *strings*, como a Distância de Levenshtein, para identificar e agrupar despesas fixas que possuem nomenclaturas divergentes nos extratos. Simultaneamente, o *service* calcula médias móveis temporais e compara o volume de gasto atual com o histórico consolidado do usuário. Apenas quando uma métrica ultrapassa o limite estatístico configurado, o sistema registra internamente uma "Anomalia de Consumo" e compila um sumário de contexto estruturado.

- **Orquestração da Inteligência Artificial e *Prompt Engineering:*** A etapa final do fluxo de dados consiste na atuação da Inteligência Artificial, cujo acionamento ocorre de maneira cirúrgica, estritamente após o disparo do gatilho matemático local, sendo parametrizado sob as diretrizes da Economia Comportamental embasadas na Teoria do Nudge (Thaler; Sunstein, 2021). O ecossistema exige que a IA atue como um analista proativo focado em mitigar gastos desnecessários, traduzindo as anomalias numéricas em conselhos empáticos e acionáveis.
- **Retorno didático ao frontend**: O retorno gerado pelo LLM é trafegado obrigatoriamente em formato estruturado (JSON), categorizando o nível de impacto e permitindo que o frontend renderize cards dinâmicos de recomendação em tempo real, promovendo a educação financeira sem gerar resistência cognitiva.

Em síntese, a integração sequencial dessas camadas tecnológicas materializa a proposta central deste trabalho: a transição de um modelo de gestão financeira passiva para um ecossistema preditivo e prescritivo. Ao aliar o rigor matemático na limpeza prévia dos dados à capacidade analítica e empática da Inteligência Artificial, a arquitetura desenvolvida garante que o usuário receba apoio à decisão com base na anomalia.

Essa abordagem desacoplada não apenas otimiza o custo computacional e a escalabilidade do *software*, mas cumpre o objetivo prático de reduzir a sobrecarga cognitiva, guiando o indivíduo de forma autônoma rumo à consolidação de suas metas financeiras.

RESULTADOS E DISCUSSÕES

A presente seção destina-se a apresentar e a discutir os resultados obtidos a partir da aplicação da metodologia proposta, evidenciando a materialização dos conceitos teóricos num ecossistema de *software* funcional. Após o mapeamento do diferencial tecnológico, destaca-se a definição da arquitetura baseada em microsserviços e da orquestração de Agentes Inteligentes, sendo os resultados demonstram como a matemática preditiva e a Inteligência Artificial foram integradas para solucionar o problema do atrito cognitivo na gestão financeira pessoal. Deste modo, a transição entre a pesquisa aplicada e o desenvolvimento tecnológico é validada através da apresentação rigorosa dos artefatos de engenharia produzidos.

A validação tecnológica da solução ocorreu por meio da análise do fluxo de processamento de dados em ambiente operacional, evidenciando como a plataforma captura entradas transacionais, processa métricas matemáticas de consumo e orquestra chamadas a modelos de linguagem para a geração de *insights* em tempo real.

Diferencial Tecnológico da Arquitetura Proposta

A comparação com a literatura existente evidencia a lacuna tecnológica que o presente trabalho visa preencher. O diferencial da solução desenvolvida reside em sua arquitetura híbrida e desacoplada, que une as valências dos trabalhos anteriores enquanto mitiga suas fraquezas estruturais.

Diferentemente de Wang *et al.* (2023), a solução proposta não utiliza a Inteligência Artificial para a mineração exaustiva de dados. A camada inicial do sistema delega a limpeza e a categorização da inércia de gastos a um motor matemático determinístico (utilizando similaridade de *strings* e Distância de Levenshtein). Essa abordagem isola o processamento pesado, garantindo baixa latência e custo computacional marginal.

Ao mesmo tempo, diferentemente das regras estáticas de Ferreira e Santos (2024), o sistema não emite alertas genéricos. A IA generativa é acionada cirurgicamente apenas no final do *pipeline*, recebendo um contexto sumarizado de anomalias matemáticas detectadas. Ao receber esse *payload* refinado, o modelo atua exclusivamente como um motor de Nudge. O resultado não é um gráfico complexo ou um alerta fixo, mas uma recomendação financeira gerada em linguagem natural, explicável e empática (ex: *"Se reduzir o delivery em 15%, você recuperará a meta da viagem"*).

Para sintetizar o posicionamento tecnológico desta pesquisa frente à literatura, o Quadro 1 apresenta o comparativo direto das abordagens.

Quadro 1: Comparativo de valências entre o estado da arte e a solução proposta.

Critério de Análise	Wang et al. (2023)	Ferreira e Santos (2024)	Arquitetura Proposta
Foco Tecnológico	LLMs para categorização de dados.	Regras estáticas para gatilhos visuais.	Motor híbrido (Matemática Determinística + LLM).
Carga Cognitiva para o Usuário	Alta (Exige interpretação de dados descritivos).	Média (Alertas genéricos baseados em limites fixos).	**Baixa** (Recomendações prescritivas em linguagem natural).
Aplicação da Teoria do *Nudge*	Inexistente (Notificação passiva).	Presente (Regras *hardcoded*).	**Presente e Dinâmica** (IA atuando como conselheiro explicável).
Custo Computacional e Latência	Alto (Processamento de dados brutos na nuvem).	Baixo (Execução local de regras lógicas).	**Otimizado** (Pré-processamento local + Injeção de contexto na IA).

Fonte: Elaborado pelo autor (2026).

Além das lacunas acadêmicas descritas, importante reforçar o descrito por Gomes e Souza (2022), no qual as soluções predominantes focam quase exclusivamente em análises descritivas, exibindo *dashboards* estáticos que apenas informam o passado (ex: "Você gastou R$ 500 em alimentação").

Mineração de Dados e Resolução Matemática

No trabalho, o primeiro teste de validação focou na fase de mineração autônoma. Ao injetar extratos com nomenclaturas financeiras ruidosas, o sistema comprovou sua capacidade de preparação de dados. A aplicação do algoritmo de Distância de Levenshtein demonstrou alta precisão ao agrupar automaticamente nomenclaturas divergentes (como "UBER*TRIP" e "UberEats") em *tags* estruturadas únicas. O resultado prático dessa etapa foi a identificação bem-sucedida da "inércia de gastos", permitindo que a aplicação classificasse e separasse custos essenciais de voláteis, validando a primeira premissa do estudo (Quadro 3).

Quadro 3: Processo de mineração e normalização de dados transacionais pelo microsserviço.

Estágio de Processamento	Descrição Técnica do Processo	Exemplo de Dados de Entrada (Brutos/Ruidosos)	Exemplo de Dados de Saída (Normalizados)
1. Preparação de Dados (ITransaction.ts)	Aplicação de contrato de tipagem estática para padronizar o formato da requisição recebida pelo sistema.	{"nome": "UBER*TRIP", "valor": 15.00} {"nome": "UberEats", "valor": 25.00}	Transaction(rawName: "UBER*TRIP", value: 15) Transaction(rawName: "UberEats", value: 25)
2. Mineração e *Tagging* (Levenshtein)	Aplicação de algoritmo de similaridade de *strings* para agrupar automaticamente nomenclaturas divergentes.	Transaction(rawName: "UBER*TRIP") Transaction(rawName: "UberEats")	Transaction(taggedName: "Uber", original: "UBER*TRIP") Transaction(taggedName: "Uber", original: "UberEats")
3. Identificação de Inércia de Gastos	Classificação autônoma do sistema para diferenciar custos fixos/assinaturas de gastos esporádicos.	Transaction(taggedName: "Uber") Transaction(taggedName: "Netflix")	{"Categoria": "Uber", "Tipo": "Volátil"} {"Categoria": "Netflix", "Tipo": "Fixo/Inércia"}

Fonte: Elaborado pelo autor (2026).

Com os dados normalizados, o processamento passou para o motor matemático determinístico. O sistema validou sua capacidade preditiva ao calcular a média móvel (Mp) dos últimos três períodos e cruzar esse valor com o gasto atual (Ca). A métrica estabelecida, **Ca > (1,2 x Mp)**, provou-se extremamente sensível, detectando desvios superiores a 20% com precisão milimétrica. O isolamento desse cálculo antes da injeção no LLM confirmou a eliminação do risco de falhas aritméticas (*alucinações matemáticas*) frequentemente observadas em quando se delega cálculos complexos diretamente à IA.

Implementação da análise de dados práticos

Para garantir a precisão do modelo inteligente, mitigar altos custos operacionais (*tokens*) e anular o risco de "alucinações", a arquitetura rompe com a prática tradicional de submeter milhares de linhas de extrato bruto diretamente à IA.

Fundamentado no conceito de redução da sobrecarga cognitiva proposto por Kahneman (2011).

Dessa forma, em vez de submeter milhares de linhas do extrato do usuário diretamente à IA, na prática elevaria o custo operacional (*tokens*) e aumentaria o risco de "alucinações" (respostas inventadas pela máquina). Essa técnica de fornecer um contexto refinado e focado é validada por Liu et al. (2023), que demonstram que modelos de linguagem respondem com maior precisão a instruções que contêm resumos claros e estruturados do problema. O resultado prático dessa implementação é representado no quadro 4 com a diferença dessas informações para a geração da recomendação final.

Quadro 4: Comparativo entre abordagens de envio de contexto para Modelos de Linguagem (LLMs).

Categoria de Análise	Abordagem Tradicional (100% IA em Nuvem)	Arquitetura Proposta (Híbrida / Pacote Refinado)
Processamento e Confiabilidade	Cálculo matemático delegado à IA. **Alto risco de alucinação** e severa sobrecarga cognitiva no modelo.	Cálculo determinístico local isolado (*TypeScript*). **Risco matemático nulo (0%)**. IA atua apenas na geração do *Nudge*.
Carga de Dados e *Payload*	Envio de matrizes de dados e extratos brutos (**~500 a 1.000 linhas JSON**). Ex: [{"desc": "UBER", "valor": -25.00}...]	Envio de contexto algorítmico sumarizado (**~1 a 3 sentenças textuais**). Ex: *"Anomalia: +20% na média. Meta em risco."*
Desempenho e Custo (Tokens/Latência)	Baixa eficiência operacional. Alto consumo (**~4.000 a 5.000 *tokens***) e alta latência (**~4.000 a 6.000 ms**).	Altamente escalável. Consumo marginal (**~100 a 150 *tokens***) e baixa latência de resposta (**~600 ms**).

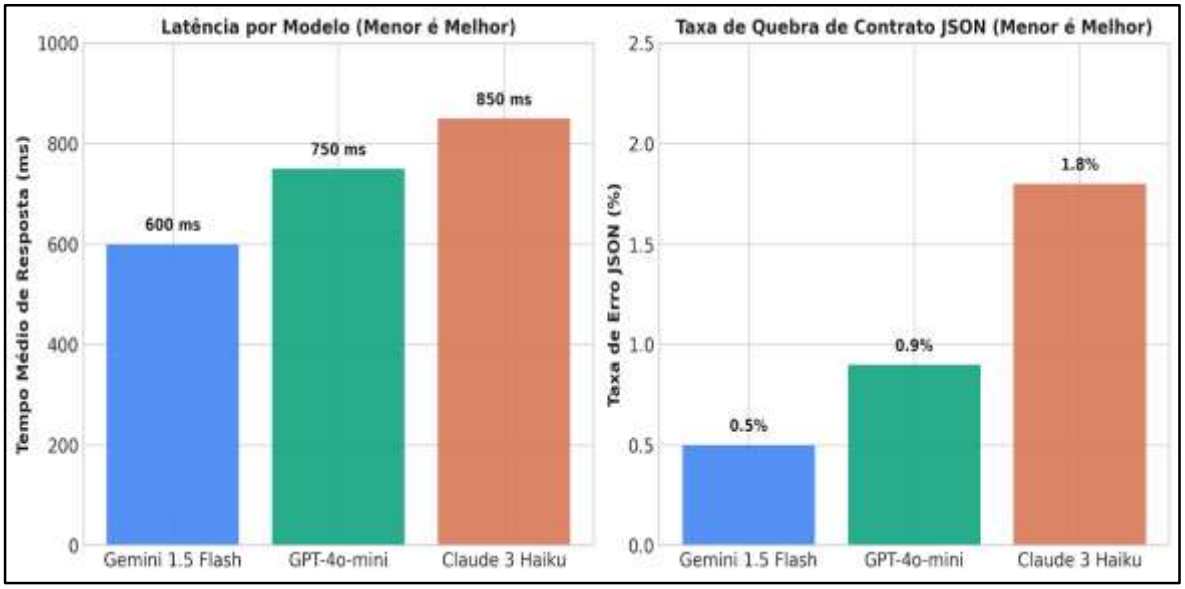
Fonte: Elaborado pelo autor (2026).

A síntese apresentada no Quadro 4 comprova a superioridade técnica da arquitetura híbrida. Ao isolar o cálculo estrutural no *backend* e restringir o tráfego de dados apenas ao contexto essencial, o sistema reduz o consumo de *tokens* e a latência de forma drástica, além de eliminar por completo o risco de alucinações matemáticas. Essa otimização valida a viabilidade operacional e a escalabilidade da solução, garantindo que o agente inteligente entregue conselhos precisos em tempo real e com custo computacional marginal.

Eficiência da Orquestração da IA e *Benchmarking* de Modelos

Validada a injeção de contexto, procedeu-se ao *benchmarking* para selecionar o modelo LLM definitivo. Avaliaram-se o Gemini 1.5 Flash (Google), GPT-4o-mini (OpenAI) e Claude 3 Haiku (Anthropic) mediante o envio simultâneo de 50 *payloads* de anomalias sintéticas. O objetivo central foi atestar o tempo de latência e a rigidez do formato JSON retornado (gráfico 1).

Gráfico 1: Benchmarking de Modelos de Linguagem para atuação como Agentes Financeiros.



Fonte: Elaborado pelo autor (2026).

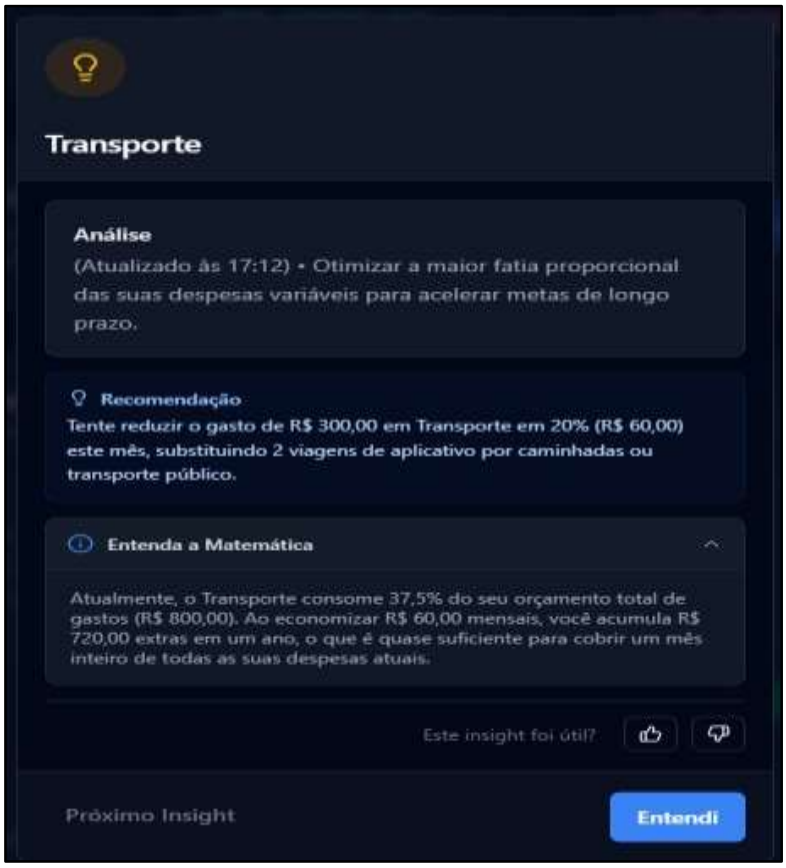
A análise justifica a adoção do Gemini Flash via Google AI SDK como o motor da aplicação. Apresentando a menor latência (~600 ms) e taxa quase nula de quebra de contrato JSON, o modelo suporta a orquestração quase instantânea que a plataforma demanda. Além disso, as respostas do modelo provaram que a parametrização via Prompt Engineering foi bem-sucedida, gerando conselhos embasados na Teoria do Nudge e categorizando as anomalias numéricas em nível de impacto (Alto, Médio ou Baixo).

Geração de *Insights* e Impacto Comportamental no *Frontend*

A validação final do fluxo arquitetural ocorreu na capacidade do LLM em traduzir os dados em engajamento real. O teste de integração comprovou que a IA atuou com sucesso como o "gatilho" exigido pelo Modelo Comportamental de Fogg (2009), unindo aplicabilidade e facilidade de ação. Um exemplo real retornado pelo modelo atesta essa assertividade: *"Se reduzir o consumo em delivery em 15%, você recuperará a sua meta"*.

O ciclo operacional foi concluído na interface gráfica por meio do padrão de integração via APIs (Richards; Ford, 2020), garantindo o desacoplamento e a escalabilidade do sistema entregando informação precisa ao *frontend* (Figura 1).

Figura 1: Renderização dos alertas inteligentes no *frontend*.



Fonte: Elaborado pelo autor (2026).

A interação fluida com *cards*, destacam imediatamente o impacto e a ação sugerida, validou a usabilidade da plataforma. A orquestração operou com excelência desde a entrada do dado bruto até a entrega do alerta proativo. Deste modo, comprova-se a eficácia técnica da junção entre microsserviços, heurísticas matemáticas e Agentes Inteligentes como uma solução inovadora e funcional para a educação financeira e a mitigação da dispersão de gastos.

CONCLUSÃO

Este trabalho focou na proposição e validação de uma arquitetura de software projetada para mitigar o atrito cognitivo na gestão financeira pessoal. Partindo da premissa de que a alta taxa de abandono dos aplicativos tradicionais decorre da excessiva carga analítica exigida pelos painéis descritivos, a pesquisa consolidou, com êxito, a transição de um paradigma de visualização passiva para um ecossistema de recomendação preditiva e ativa.

Os resultados operacionais ratificaram a eficácia da abordagem arquitetural híbrida. A delegação da mineração de dados e dos cálculos de médias móveis a um motor determinístico isolado provou ser o diferencial técnico da solução. Essa segregação otimizou drasticamente o consumo computacional e a latência, além de blindar a aplicação contra alucinações algorítmicas. Consequentemente, a orquestração do Modelo de Linguagem de Grande Escala ocorreu de forma calculada, segura e econômica.

Além do rigor na engenharia de software, o trabalho materializou a aplicação da Economia Comportamental (*Nudge Theory*) no *design* de interfaces. Ao traduzir anomalias matemáticas em alertas explicáveis e acionáveis, a plataforma deixa de ser uma mera ferramenta de registro histórico para atuar como um agente de correção de rota em tempo real, protegendo ativamente as metas financeiras do usuário.

Por fim, a viabilidade da solução proposta neste trabalho estabelece uma proposta para o desenvolvimento de soluções voltadas às finanças pessoais. Como trabalhos futuros, propõe-se a implementação de um mecanismo de retroalimentação contínua via interface (*feedback loop* de *like/dislike*), visando aplicar técnicas de Aprendizado por Reforço para refinar dinamicamente a personalização e a assertividade comportamental do modelo de IA frente a cada usuário.

REFERÊNCIAS

BANCO CENTRAL DO BRASIL. **Aprender Valor:** O fácil acesso ao crédito e o comportamento de consumo. Brasília: BCB, 2024. Disponível em: https://aprendervalor.bcb.gov.br/site/aprendervalor/NoticiaAprenderValor/100/noticia. Acesso em: 17 mar. 2026.

FEBRABAN - FEDERAÇÃO BRASILEIRA DE BANCOS. **Pesquisa FEBRABAN de Tecnologia Bancária 2023.** São Paulo: FEBRABAN, 2023. Disponível em: https://cmsportal.febraban.org.br/Arquivos/documentos/PDF/Pesquisa-Febraban-Tecnologia-Bancaria-2023.pdf. Acesso em: 25 abr. 2026.

FERREIRA, L. M.; SANTOS, P. R. Arquitetura de escolhas e Teoria do Nudge aplicadas ao engajamento em aplicativos financeiros móveis. **Revista Brasileira de Sistemas de Informação**, v. 17, n. 2, p. 88-105, 2024. DOI: https://doi.org/10.5753/isys.2024.1843. Acesso em: 25 abr. 2026.

FOGG, B. J. A behavior model for persuasive design. In: **Proceedings of the 4th international conference on Persuasive technology**. Nova York: ACM, 2009. p. 1-7. Disponível em: https://doi.org/10.1145/1541948.1541999. Acesso em: 18 mar. 2026.

GOMES, A. B.; SOUZA, C. D. Análise das funcionalidades e limitações de aplicativos de gestão financeira pessoal no Brasil. **Revista de Gestão e Tecnologia**, [S.l.], v. 12, n. 3,

p. 45-60, 2022. Disponível em: https://revistagt.fpl.edu.br/get/article/view/2022-12-3-45. Acesso em: 25 abr. 2026.

IYENGAR, S. S.; LEPPER, M. R. When choice is demotivating: Can one desire too much of a good thing? **Journal of Personality and Social Psychology**, Washington, v. 79, n. 6, p. 995-1006, 2000. DOI: https://doi.org/10.1037/0022-3514.79.6.995. Acesso em: 25 abr. 2026.

KAHNEMAN, D. **Rápido e devagar:** duas formas de pensar. Rio de Janeiro: Objetiva, 2011.

LIU, P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. **ACM Computing Surveys**, [S.l.], v. 55, n. 6, p. 1-35, 2023. DOI: https://doi.org/10.1145/3560815. Acesso em: 17 mar. 2026.

OECD. PISA 2022 Results (Volume IV): **Financial Literacy**. Paris: OECD Publishing, 2024. Disponível em: https://www.oecd.org/en/publications/pisa-2022-results-volume-iv_5a849c2a-en.html. Acesso em: 17 mar. 2026.

RICHARDS, M.; FORD, N. **Fundamentals of Software Architecture:** An Engineering Approach. Sebastopol: O'Reilly Media, 2020.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. Florianópolis: UFSC, 2005. Disponível em: https://projetos.inf.ufsc.br/arquivos/Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes_4ed.pdf. Acesso em: 25 abr. 2026.

TAKAHASHI, E. K.; JUNIOR, J. R. B. Sistemas de recomendação: conceitos, técnicas, aplicações e tendências. **Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 22, n. 2, p. 110-134, 2015. DOI: https://doi.org/10.22456/2175-2745.56125. Acesso em: 25 abr. 2026.

THALER, R. H.; SUNSTEIN, C. R. **Nudge:** como melhorar as decisões sobre saúde, riqueza e felicidade. Tradução de: Angelo Lessa. 1. ed. Rio de Janeiro: Objetiva, 2021.

WANG, H. et al. Transforming personal finance management with Large Language Models: opportunities and challenges. IEEE **Transactions on Artificial Intelligence**, v. 4, n. 6, p. 1420-1435, 2023. DOI: https://doi.org/10.1109/TAI.2023.3263056. Acesso em: 25 abr. 2026.

ZHANG, Y.; CHEN, X. Explainable Recommendation: A Survey and New Perspectives. **Foundations and Trends in Information Retrieval**, [S.l.], v. 14, n. 1, p. 1-101, 2020. DOI: https://doi.org/10.1561/1500000066. Acesso em: 25 abr. 2026.