UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Curso de Bacharelado em Engenharia de Computação



Trabalho de Conclusão de Curso

Uso de Agentes Generativos como Ferramenta de Suporte à Decisão no Varejo

Bernardo Garcia Ferrão

Pelotas, 2026

Bernardo Garcia Ferrão

Uso de Agentes Generativos como Ferramenta de Suporte à Decisão no Varejo

Trabalho de Conclusão de Curso apresentado ao Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Engenharia de Computação.

Orientadora: Profa. Dra. Míriam Blank Born

Pelotas, 2026

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

F369u Ferrão, Bernardo Garcia

Uso de agentes generativos como ferramenta de suporte à decisão no varejo [recurso eletrônico] / Bernardo Garcia Ferrão ; Míriam Blank Born, orientadora. — Pelotas, 2026.
97 f. : il.

Trabalho de Conclusão de Curso (Graduação) - Engenharia de Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2026.

1. Simulação. 2. Sistemas multiagente. 3. LLM. 4. Agentes generativos. 5. Varejo. I. Born, Míriam Blank, orient. II. Título.

CDD 005

Elaborada por Simone Godinho Maisonave CRB: 10/1733

Bernardo Garcia Ferrão

Uso de Agentes Generativos como Ferramenta de Suporte à Decisão no Varejo

Trabalho de Conclusão de Curso aprovado, como requisito parcial, para obtenção do grau de Bacharel em Engenharia de Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 17 de julho de 2026

Banca Examinadora:

Profa. Dra. Míriam Blank Born (orientadora)
Doutora em Computação pela Universidade Federal de Pelotas.

Prof. Marilton Sanchotene de Aguiar.
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Bruno Cascaes Alves
Mestre Ciência da Computação pela Universidade Federal de Pelotas.

Dedico este trabalho aos meus pais e avós...

AGRADECIMENTOS

A conclusão deste curso e a elaboração deste trabalho não seriam possíveis sem o auxílio de pessoas essenciais na minha trajetória. Agradeço imensamente aos meus pais e avós, que foram meu porto seguro, proporcionando apoio incondicional de diversas formas.

À minha namorada, Amanda Sampaio Laguna, pelo incentivo contínuo, por me amparar nos momentos difíceis e me impulsionar a evoluir sempre.

Ao meu primo, Rafael Seabra Ferrão, pela amizade e parceria constantes durante nossa jornada morando juntos em Pelotas.

Aos colegas que viraram grandes amigos: Ana, Mathias, Vanessa, Josué, Rafael, Jean, Pedro e Santiago, pela troca de conhecimentos e experiências tanto em sala de aula quanto fora dela.

Aos meus amigos de Camaquã do grupo Esquenta: Caio, Cássio, Guilherme, Alexandre e Henrique pela união e irmandade de sempre, independentemente do tempo sem nos vermos.

Finalmente, expresso minha gratidão à minha orientadora, Profa. Míriam, pela compreensão, por ter acreditado neste projeto e por todo o suporte técnico e acadêmico oferecido.

Excelência não é um ato, mas um hábito. Somos o que fazemos repetidamente.
— ARISTÓTELES

RESUMO

FERRÃO, Bernardo Garcia. **Uso de Agentes Generativos como Ferramenta de Suporte à Decisão no Varejo**. Orientadora: Míriam Blank Born. 2026. 97 f. Trabalho de Conclusão de Curso (Engenharia de Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2026.

O varejo é um setor altamente competitivo, que demanda a contínua experimentação de novas estratégias de gestão. No entanto, testá-las no mundo real custa caro e envolve riscos operacionais, o que muitas vezes as torna inviáveis. A modelagem de Sistemas Multiagente (SMA), abordagem *bottom-up* que captura comportamentos emergentes, é comumente usada para sistemas complexos. A evolução dos agentes generativos, viabilizados pelos modelos de linguagem de grande escala (LLMs) e dotados de memória, planejamento e interação social realista, amplia essa possibilidade: o SMA oferece o ambiente e as regras, enquanto os LLMs assumem a camada cognitiva, modelando a complexidade do comportamento do consumidor no varejo. O objetivo principal deste trabalho é investigar o uso de agentes generativos como ferramenta de suporte à decisão no varejo, avaliando se a camada cognitiva baseada em LLM representa o comportamento de uma loja real de forma mais fiel do que uma abordagem clássica baseada em regras. Uma loja é modelada como um sistema multiagente e simulada sob as duas abordagens, cujos indicadores emergentes são confrontados com os dados reais do estabelecimento; sobre o modelo validado, simulam-se estratégias contrafactuais de *layout*, promoção e atendimento em um ambiente seguro (*sandbox*), sem risco operacional. A principal contribuição para a Computação é a investigação e a aplicação de agentes generativos em Sistemas Multiagente para representar o comportamento do consumidor no varejo. Os resultados mostram que a melhor abordagem depende da análise pretendida: para leituras mecânicas, como a variação dos indicadores sob diferentes fluxos de clientes ou a configuração padrão da loja, a abordagem clássica é preferível, por dispensar a captação de contexto e não ter custo; já a generativa se destaca nos cenários contrafactuais dependentes de contexto, reagindo de forma plausível a promoções, mudanças de *layout* e diferentes tons de atendimento. A principal limitação observada é a ausência de compras em grupo e de dinâmicas familiares, pois os clientes são modelados como indivíduos isolados, o que penaliza sobretudo a representação do departamento infantil. Como trabalhos futuros, propõem-se a modelagem de núcleos familiares e a avaliação de modelos de LLM locais, para reduzir latência e custo. Conclui-se que as duas abordagens são complementares, e seu uso combinado constitui uma ferramenta de apoio à decisão do tipo *sandbox*.

Palavras-chave: simulação; sistemas multiagente; LLM; agentes generativos; varejo.

ABSTRACT

FERRÃO, Bernardo Garcia. **Use of Generative Agents as a Decision Support Tool in Retail**. Advisor: Míriam Blank Born. 2026. 97 f. Undergraduate Thesis (Computer Engineering) – Technology Development Center, Federal University of Pelotas, Pelotas, 2026.

Retail is a highly competitive sector that demands continuous experimentation with new management strategies. However, testing them in the real world is expensive and entails operational risks, which often makes it unfeasible. Multi-Agent Systems (MAS) modeling, a bottom-up approach that captures emergent behavior, is commonly used for complex systems. The evolution of generative agents, enabled by Large Language Models (LLMs) and endowed with memory, planning, and realistic social interaction, broadens this possibility: the MAS provides the environment and the rules, while the LLMs handle the agents' cognitive layer, modeling the complexity of consumer behavior in retail. The main goal of this work is to investigate the use of generative agents as a decision-support tool in retail, assessing whether the LLM-based cognitive layer represents the behavior of a real store more faithfully than a classic, rule-based approach. A store is modeled as a multi-agent system and simulated under both approaches, whose emergent indicators are confronted with real data from the establishment; on the validated model, counterfactual strategies of layout, promotion, and customer service are simulated in a safe environment (sandbox), without operational risk. The main contribution to Computer Science is the investigation and application of generative agents in Multi-Agent Systems to represent consumer behavior in retail. The results show that the most suitable approach depends on the intended analysis: for mechanical readings, such as how the indicators change under different customer flows or the store's default configuration, the classic approach is preferable, as it requires no context capture and has no cost; the generative approach, in turn, stands out in context-dependent counterfactual scenarios, reacting plausibly to promotions, layout changes, and different service tones. The main limitation observed is the absence of group purchases and family dynamics, as customers are modeled as isolated individuals, which especially penalizes the representation of the children's department. As future work, we propose modeling family units and evaluating local LLM models to reduce latency and cost. We conclude that the two approaches are complementary, and their combined use constitutes a sandbox-type decision-support tool.

Keywords: simulation; multi-agent systems; LLM; generative agents; retail.

LISTA DE FIGURAS

Figura 1 Visual da plataforma GAMA 29
Figura 2 Visual do framework AGNO (AgentOS) 30

Figura 3 Arquitetura do projeto: Motores espacial e cognitivo; e ponte. 44
Figura 4 Planta arquitetônica 45
Figura 5 Grafo, arestas e pontos de avaliação 47
Figura 6 Funil de compra do agente cliente. 55
Figura 7 Hierarquia de classes dos agentes. 56
Figura 8 Funil de ações do agente vendedor. 57

Figura 9 Mapa de calor de ocupação da loja pelos compradores: clássico (esquerda) e generativo (direita). 70
Figura 10 Diferença de ocupação entre as abordagens (generativo menos clássico), apenas compradores. 70
Figura 11 Exemplo de alucinação do vendedor: o item e o preço oferecidos não constavam no catálogo. 72
Figura 12 Justificativas de navegação relacionadas ao fluxo de pessoas, com a percepção de layout ligada. 76
Figura 13 Mapa de calor de ocupação pelos compradores, média das *seeds* 1, 7 e 42: *baseline* (esquerda) e com o Infantil ampliado (direita). . . 77
Figura 14 Mapa de calor das duas abordagens sob o Infantil ampliado 78
Figura 15 Exemplo de conversa sobre a promoção nos calçados. 79
Figura 16 Exemplo do tom neutro: o vendedor informa as opções pedidas, sem elogiar nem pressionar. 80
Figura 17 Exemplo do tom incisivo: o vendedor insiste com urgência (escassez) sobre a mesma peça, e o cliente recusa repetidamente, sem fechar a compra. 81
Figura 18 Exemplo do Tom calmo: o vendedor observa a cesta e sugere, sem pressão, uma peça que combina. 81
Figura 19 Exemplo do tom caloroso: o vendedor elogia a escolha e cria conexão emocional. 82

LISTA DE TABELAS

Tabela 1 Comparativo entre os trabalhos. 38

Tabela 2 Fontes de dados que abastecem o modelo. 40
Tabela 3 Análise de sensibilidade das premissas 41
Tabela 4 Origem das faixas de variação de cada parâmetro 41

Tabela 5 Áreas das seções: modelo versus loja real. 45
Tabela 6 Departamento de cada família pelo prefixo do código 48
Tabela 7 Entrantes por faixa de horário e dia da semana (abril/2026) 49
Tabela 8 Janelas de chegada: intervalo, volume e peso de cada janela no total do dia 49
Tabela 9 Origem das características da persona do cliente. 51
Tabela 10 Fatores usados no cálculo do orçamento. 52
Tabela 11 Despesa com vestuário por classe de renda (POF 2018, recorte Sul; valores per capita em reais de abril de 2026). 53
Tabela 12 Renda per capita e fração de vestuário por décimo (Região Sul, abril de 2026). 53
Tabela 13 Contexto-base enviado à LLM em toda decisão. 61

Tabela 14 Tempo de execução e custo por run no modo de pico. 64
Tabela 15 Registro cognitivo armazenado no transcript 65
Tabela 16 Arquivos CSV gerados pelo GAMA com o desfecho da simulação . 65
Tabela 17 Comparação dos KPIs agregados entre abordagens e a realidade. . 66
Tabela 18 Análise de inversão: valor que cada premissa precisaria assumir, isoladamente, para o clássico atingir a cesta real (2,59 itens por cupom). 67
Tabela 19 Mix por departamento (% do valor vendido). 67
Tabela 20 Itens do top-N por quantidade de cada abordagem que constam no top-N real (também por quantidade de unidades vendidas). 69
Tabela 21 Dez itens mais comprados por unidade em cada abordagem (soma das 8 *seeds*). 69
Tabela 22 Padrão de visitas, conversão por visita e mix de valor por departamento (abordagem generativa). 71
Tabela 23 Padrão de visitas, conversão por visita e mix de valor por departamento (abordagem clássica). 71
Tabela 24 Faixas de ocupação e o rótulo qualitativo de lotação percebido pelo agente. 75

Tabela 25 Percentual de decisões de navegação que citam a lotação/fluxo na justificativa. 76
Tabela 26 Efeito do redimensionamento (Infantil ampliado) sobre a participação do Infantil no mix, por abordagem 77
Tabela 27 Efeito da promoção sobre a participação do departamento no valor vendido, o ticket médio e a cesta 79
Tabela 28 Texto injetado no *prompt* da LLM indicando o tom a ser usado. . . . 80
Tabela 29 Efeito do tom de abordagem do vendedor sobre os indicadores. . . 82

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
API	Application Programming Interface
B2C	Business to Consumer
BFS	Breadth-First Search (busca em largura)
CNC	Confederação Nacional do Comércio de Bens, Serviços e Turismo
CSV	Comma-Separated Values (valores separados por vírgula)
FIFO	First In, First Out (primeiro a entrar, primeiro a sair)
GAML	GAMA Modeling Language (linguagem do software GAMA)
GUI	Graphical User Interface (interface gráfica do usuário)
HTTP	Hypertext Transfer Protocol
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
ICMS	Imposto sobre Circulação de Mercadorias e Serviços
IPCA	Índice Nacional de Preços ao Consumidor Amplo
ISS	Imposto Sobre Serviços
JSON	JavaScript Object Notation
JSONL	JSON Lines
KPI	Key Performance Indicator (indicador-chave de desempenho)
LLM	Large Language Models (grandes modelos de linguagem)
MSBA	Modelagem e Simulação Baseada em Agentes
OAT	One-Factor-At-a-Time (análise um fator por vez)
PAARS	Persona Aligned Agentic Retail Shoppers
PAC	Pesquisa Anual de Comércio
PIB	Produto Interno Bruto
PNAD	Pesquisa Nacional por Amostra de Domicílios
POF	Pesquisa de Orçamentos Familiares

RAG	Retrieval-Augmented Generation (geração aumentada por recuperação)
REST	Representational State Transfer
RS	Rio Grande do Sul
SBVC	Sociedade Brasileira de Varejo e Consumo
SDK	Software Development Kit (kit de desenvolvimento de software)
SKU	Stock Keeping Unit (unidade de manutenção de estoque)
SMA	Sistemas Multiagente

SUMÁRIO

1 INTRODUÇÃO 18

2 REFERENCIAL TEÓRICO 21
2.1 O Setor Varejista 21
2.1.1 Definição e Panorama Econômico 21
2.1.2 Barreiras Estruturais e a Volatilidade do Mercado Brasileiro 22
2.1.3 A Ciência do Consumo e o Comportamento do Cliente 22
2.1.4 Adaptabilidade e Simulação 22
2.2 Sistemas Multiagente (SMA) 23
2.2.1 Fundamentos de modelagem e simulação 23
2.2.2 Evolução e Conceitos Fundamentais 23
2.2.3 Modelagem Bottom-up e Comportamento Emergente 24
2.2.4 Limitações dos Agentes Clássicos 24
2.3 Inteligência Artificial e Grandes Modelos de Linguagem (LLMs) . . . 24
2.3.1 IA Simbólica 24
2.3.2 IA Conectivista 25
2.3.3 IA Moderna 25
2.4 Agentes Generativos 26
2.4.1 Definição e Paradigma 26
2.4.2 Arquitetura Cognitiva: Memória, Reflexão e Planejamento 26
2.4.3 Engenharia de Prompt 27

3 REFERENCIAL TECNOLÓGICO 28
3.1 GAMA 28
3.2 Agno 29
3.3 API Agentic MAS 30

4 TRABALHOS RELACIONADOS 31
4.1 Agentes generativos 31
4.1.1 Park et al. (2023) — Generative Agents: Interactive Simulacra of Human Behavior 31
4.1.2 Kato et al. (2024) — Constructing LLM-based Agent Model and Its Application to Multi-Agent Simulation 32
4.2 Generativos aplicados ao consumo e ao varejo 33
4.2.1 Mansour et al. (2025) — PAARS: Persona Aligned Agentic Retail Shoppers 33
4.2.2 Li, Wei & Wang (2025) — Predicting Behaviors with Large Language Model (LLM)-Powered 34

4.2.3 Oliveira (2025) — Desenvolvimento de um sistema multiagente baseado em inteligências artificiais generativas para avaliação de qualidade em atendimento ao cliente . . . 35
4.3 Dinâmica espacial e movimentação de consumidores . . . 35
4.3.1 Horikomi & Mizuno (2025) — Interaction-aware agent-based simulation of customer trajectories in retail stores with transformer architectures . . 35
4.3.2 Pantano et al. (2021) — Enhancing store layout decision with agent-based simulations of consumers' density . . . 36
4.4 Posicionamento do Trabalho Proposto . . . 37

5 METODOLOGIA . . . 39
5.1 Visão geral . . . 39
5.2 Estudo de caso . . . 39
5.3 Modelagem inicial . . . 40
5.4 Calibração e validação . . . 42
5.5 Desenho experimental e métricas . . . 43

6 DESENVOLVIMENTO . . . 44
6.1 Ambiente, layout e zonas . . . 44
6.2 Movimentação e navegação . . . 46
6.3 Catálogo de produtos . . . 47
6.4 Chegada e volume . . . 48
6.5 Agente Cliente . . . 50
6.5.1 Persona . . . 50
6.5.2 *Buyer* ou *Browser* . . . 51
6.5.3 Orçamento . . . 51
6.5.4 Funil de compra . . . 53
6.6 Agente Vendedor . . . 55
6.7 Filas e atendimento . . . 57
6.8 LLM e cognição . . . 58
6.8.1 Arquitetura da ponte GAMA–Agno . . . 58
6.8.2 As decisões cognitivas e seu contexto . . . 59
6.8.3 Persona determinística e memória . . . 61

7 ANÁLISE DOS RESULTADOS . . . 63
7.1 Custo, latência e robustez . . . 63
7.2 Coleta de dados . . . 64
7.3 Experimento principal . . . 65
7.3.1 Conversão, Ticket, Cesta e Faturamento . . . 66
7.3.2 Mix por departamento . . . 67
7.3.3 Ranking de itens . . . 68
7.3.4 Dinâmica espacial . . . 69
7.3.5 Análise qualitativa das decisões . . . 72
7.4 Experimentos contrafactuais . . . 74
7.4.1 Layout . . . 75
7.4.2 Promoção . . . 78
7.4.3 Atendimento . . . 80
7.5 Discussão dos resultados . . . 83

8 CONCLUSÃO 85
8.1 Limitações 86
8.2 Trabalhos futuros 87

REFERÊNCIAS 89

APÊNDICE A TEMPLATES DE PROMPT 94
A.1 Description Cliente 94
A.2 História 94
A.3 Navegação 95
A.4 Escolha de item 95
A.5 Conversa (lado do cliente) 96
A.6 Description Vendedor 96
A.7 Conversa (lado do vendedor) 97

1 INTRODUÇÃO

O varejo, hoje um setor de grande importância para a economia brasileira e mundial, possui origem milenar, com suas raízes datadas entre 9.000 e 6.000 a.C. Inicialmente, surgiu com a prática do escambo, que se trata da troca direta sem o uso de dinheiro. Posteriormente, por volta de 3.000 a.C., essa prática se estruturou com a utilização das primeiras formas de dinheiro. No cenário brasileiro, a atividade varejista remonta ao período colonial, época em que os produtos naturais cultivados, tais como café, cana-de-açúcar e algodão, além de exportados, eram adquiridos pelos moradores (Varotto, 2006; Parente; Barki, 2014).

O varejo modernizou-se com o passar do tempo, influenciado pela globalização e também pela evolução tecnológica. Primeiro, os clientes passaram a escolher seus produtos no mostruário, não mais diretamente no balcão com o vendedor: o chamado autosserviço (Parente; Barki, 2014). Hoje, com a internet, a área conhecida como *e-commerce*, responsável pelas vendas *online*, ocupa cerca de 20% das vendas mundiais (eMarketer, 2024). Entretanto, apesar da conveniência digital, o ambiente físico permanece central e cada vez mais integrado aos canais digitais, em uma jornada de compra *omnichannel* na qual o consumidor transita de forma fluida entre os pontos de contato físicos e digitais (Deloitte, 2024), o que mostra que, para se manter modernizada, a marca deve ser multiplataforma.

Essa área, que é um dos pilares da economia nacional, representava cerca de 25,23% do PIB brasileiro em 2023 (SBVC, 2024). O ambiente é dinâmico, altamente competitivo e precisa estar sempre se atualizando. Além disso, os comércios brasileiros enfrentam barreiras estruturais, como altas taxas, que atrasam o avanço da área (Dos santos; Seabra, 2025; SBVC, 2024). Em 2026, essa situação é agravada pela grande velocidade com que o mercado se move e pela necessidade de respostas rápidas a mudanças no comportamento de consumo.

A modernização de técnicas e abordagens de venda é essencial para a resiliência organizacional em mercados voláteis (Vial, 2019). Entretanto, há receio na sua aplicação devido à incerteza de sucesso, visto que este depende principalmente de variáveis comportamentais humanas, que são complexas e de difícil compreensão em modelos

tradicionais. Um erro estratégico pode gerar problemas na receita, na imagem e na insatisfação dos clientes com a marca (Keller, 2020). Nesse setor, a margem para erros é mínima; a carência de um *sandbox*, que é um ambiente de testes virtual e seguro, torna a inovação um processo de incertezas e alto custo. Nesse cenário, é evidenciada a necessidade de ferramentas que consigam simular o resultado da aplicação de novas técnicas antes de sua implementação fixa, considerando a subjetividade das decisões humanas.

Para solucionar essa lacuna e permitir a experimentação estratégica em um ambiente seguro, este projeto propõe o desenvolvimento de uma ferramenta de suporte à decisão. Essa ferramenta é composta pelo uso de sistemas multiagente (SMA) para gerenciar as ações das entidades no ambiente físico e de inteligência artificial generativa para emulação cognitiva. Diferentemente de agentes clássicos, que são baseados em regras estáticas, essa solução utiliza grandes modelos de linguagem (LLM, do inglês *Large Language Models*) e engenharia de *prompt* para dar origem aos agentes generativos; esses agentes, dotados de memória, reflexão e planejamento, conseguem mimetizar o comportamento humano com mais profundidade (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023).

A implementação técnica da ferramenta emprega duas plataformas: a ferramenta GAMA é responsável pelo motor espacial, gerindo o ambiente da loja e as ações dos agentes nele (Taillandier; Gaudou; Grignard; Huynh; Marilleau; Caillou; Philippon; Drogoul, 2019); já o *framework* Agno é responsável por dar vida aos agentes generativos. Nele são gerenciados os perfis de personalidade dos agentes, por meio de engenharia de *prompt*, e o processamento do raciocínio, com memória e LLM, retornando ao GAMA o planejamento das decisões (Agno, 2024). A comunicação entre as plataformas é realizada por meio do protocolo HTTP, com as mensagens estruturadas em formato JSON. Essa estratégia procura capturar a subjetividade do comportamento humano a fim de simular a jornada do consumidor moderno com maior fidelidade.

O objetivo principal deste trabalho é investigar o uso de agentes generativos como ferramenta de suporte à decisão no varejo, avaliando se a camada cognitiva baseada em LLM representa o comportamento de uma loja real de forma mais fiel do que uma abordagem clássica, baseada em regras. A hipótese central deste trabalho é a de que a integração de uma camada cognitiva baseada em LLMs a um sistema multiagente melhora a capacidade de mimetização do comportamento do consumidor em ambiente de loja, já que esses comportamentos são dependentes de contexto e não pré-planejados, sendo uma alternativa mais fiel do que a baseada em regras. Para alcançar esse objetivo, foram definidos os seguintes objetivos específicos:

- modelar uma loja real como um sistema multiagente, a partir de dados empíricos;
- simular a loja sob as duas abordagens, a clássica e a generativa;

- validar o modelo, confrontando os indicadores emergentes com os dados reais do estabelecimento;
- simular estratégias contrafactuais de *layout*, promoção e atendimento, avaliando a plausibilidade das reações de cada abordagem.

A simulação monitora KPIs financeiros e operacionais: conversão de venda e *ticket* médio, somados a indicadores de comportamento espacial baseados na ciência do *shopping*. Entre eles, o sistema mede o *dwell time* (tempo de permanência) por setor e na loja como um todo, além da taxa de interceptação de vendedores, utilizada como diagnóstico do atendimento. Conceitos correlacionados descritos por (Underhill, 2009), como o *butt-brush effect*, a zona de descompressão e o conforto ambiental, auxiliaram na modelagem, mas não são quantificados. A ferramenta busca representar as interações sociais e as dinâmicas de decisão sob o princípio da racionalidade limitada (do inglês, *bounded rationality*) (Simon, 1955), o qual propõe que os seres humanos não buscam a decisão perfeita, mas sim a satisfatória. Isso se deve à lacuna de informações e às limitações cognitivas e de tempo. Ao reproduzir o comportamento humano de forma plausível, a ferramenta permite a realização de testes A/B agênticos (Mansour; Perelli; Mainetti; Davidson; D'amato, 2025), possibilitando a previsão de resultados de diferentes estratégias e abordagens de gestão (Li; Wei; Wang, 2025), consolidando essa simulação como uma ferramenta de suporte à decisão promissora.

As principais contribuições deste trabalho para a Computação são: o uso de agentes generativos em sistemas multiagente no domínio comercial do varejo físico; a comparação de representação de loja e comportamento do consumidor entre as abordagens clássica e generativa, confrontadas com os dados reais de um estabelecimento; e a delimitação dos limites e das potencialidades da camada cognitiva, apontando o uso combinado das duas abordagens como ferramenta de apoio à decisão do tipo *sandbox*.

Esta monografia está estruturada da seguinte maneira: o Capítulo 2 é o referencial teórico, que aborda a definição e a contextualização dos conceitos presentes neste trabalho. O Capítulo 3 apresenta o referencial tecnológico, com uma breve descrição das ferramentas utilizadas. No Capítulo 4 são discutidos os trabalhos relacionados. A metodologia se encontra no Capítulo 5, retratando como foi realizada a simulação. O Capítulo 6 detalha o desenvolvimento da simulação. No Capítulo 7 é realizada a análise dos resultados obtidos, incluindo a discussão. Por fim, o Capítulo 8 traz a conclusão da pesquisa, com as limitações e os trabalhos futuros; na sequência, o apêndice reúne os templates de prompt.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta a fundamentação teórica que sustentou o desenvolvimento deste trabalho. A estrutura divide-se em quatro eixos principais: o setor varejista, que constitui o estudo de caso; a modelagem e simulação multiagente, adotada como abordagem técnica de desenvolvimento; a inteligência artificial, campo no qual este estudo se insere; e os agentes generativos, paradigma utilizado como base para esta pesquisa.

2.1 O Setor Varejista

Esta seção caracteriza o setor varejista, que constitui o estudo de caso deste trabalho. Primeiro, aborda-se o panorama econômico do setor e as barreiras estruturais que o tornam um ambiente de alto risco; logo depois, examina-se a ciência do consumo, que dirige o comportamento do cliente, e, por fim, a simulação como possível ferramenta para testar hipóteses sem risco operacional.

2.1.1 Definição e Panorama Econômico

O setor varejista é caracterizado pela comercialização direta de bens ou serviços ao consumidor final (B2C) (do inglês - *Business-to-Consumer*). O varejo age como elo final na cadeia de suprimentos, sendo o ponto onde a oferta encontra a demanda. No contexto brasileiro, a vasta capilaridade do varejo se faz presente no cotidiano do brasileiro, abrangendo desde a compra de itens de consumo básico em supermercados até segmentos especializados, como moda e saúde. Essa diversidade ocorre apesar de a legislação brasileira estabelecer distinções tributárias entre a venda de mercadorias (incidência de ICMS) e a prestação de serviços (incidência de ISS).

Como pilar da economia nacional, o varejo é suporte de uma grande parcela do produto interno bruto (PIB) e da geração de empregos. No ano de 2023, esse setor representava aproximadamente 25,23% do PIB nacional (SBVC, 2024). Além da grande influência na economia do país, a importância do varejo também se faz em sua capacidade de empregabilidade; o setor corresponde a cerca de 72,7% das vagas de

comércio formal (IBGE, 2025), totalizando 7,7 milhões de trabalhadores formais no setor varejista.

2.1.2 Barreiras Estruturais e a Volatilidade do Mercado Brasileiro

No entanto, o crescimento desse setor no Brasil enfrenta algumas barreiras estruturais: a complexa carga tributária brasileira, aliada à necessidade de constante modernização de abordagens, bem como a burocracia em excesso influenciam o fechamento de empresas desse setor, limitam grandes empresas e, de forma mais crítica, acabam tornando impossível a sobrevivência de pequenos empreendimentos (Dos santos; Seabra, 2025). Entre janeiro e setembro de 2024, foram registradas 314,4 mil aberturas de novas lojas; porém, cerca de 189,7 mil estabelecimentos fecharam nesse mesmo período (CNC, 2024). Nesse setor, a cada 10 estabelecimentos novos que entram em vigor, 6 estão fadados ao fechamento de portas.

2.1.3 A Ciência do Consumo e o Comportamento do Cliente

Para entender a profundidade do funcionamento do varejo, podemos analisar indicadores de consumo e comportamentais dentro do ponto de venda. Segundo (Underhill, 2009), a sobrevivência de um empreendimento nesse setor está intimamente ligada à conversão de visitantes em clientes, e essa conversão depende da percepção do consumidor de variáveis ambientais e psicológicas.

Entre os indicadores fundamentais citados por Underhill, sobressai-se o *dwell time*, observando-se que, quanto maior o tempo que os clientes ficavam no estabelecimento, maior a probabilidade de consumo, influenciando diretamente no número de vendas. Ademais, outros conceitos afetam essa variável, tais como: a “zona de descompressão”, referente aos primeiros metros a partir da entrada, onde o cliente ainda está se acostumando com o ambiente e acaba não olhando produtos nessa área; a “taxa de interceptação” dos funcionários, que deve ser alta a fim de não deixar o cliente “perdido”; e o evento chamado *butt-brush*, que demonstra que os clientes não gostam de ser tocados por trás. Esse efeito está diretamente ligado à densidade de pessoas no espaço de venda: conforme a lotação aumenta, cresce a sensação de desconforto e de invasão do espaço pessoal (Underhill, 2009), o que altera a psicologia e a navegação do consumidor, podendo levá-lo a evitar áreas congestionadas, reduzir sua permanência ou abandonar a compra. Tais fatores demonstram que o layout da loja e a interação social influenciam diretamente o volume de vendas.

2.1.4 Adaptabilidade e Simulação

Diante desse cenário, a adaptação torna-se um fator de sobrevivência organizacional, uma vez que as organizações precisam adaptar continuamente suas estruturas e processos para sobreviver em ambientes dinâmicos (Vial, 2019). A busca para au-

mentar a produtividade não se limita apenas à melhoria da logística, mas estende-se à adoção de estratégias modernas de venda, especialmente aquelas baseadas no comportamento do consumidor moderno (Kotler; Kartajaya; Setiawan, 2021; Lemon; Verhoef, 2016). Contudo, a implementação de novas técnicas pode ser complexa, uma vez que a variável central desse setor, o comportamento humano, é inerentemente imprevisível. Experimentos mal-sucedidos em ambiente real podem comprometer a receita, a imagem da marca e gerar insatisfação dos clientes (Bughin; Hazan; Labaye; Manyika; Dahlström; Ramaswamy, 2019; Keller, 2020). A simulação pré-implementação é uma prática que possibilita testar hipóteses sem risco operacional, fundamentando a tomada de decisão.

2.2 Sistemas Multiagente (SMA)

Esta seção apresenta a modelagem e a simulação baseadas em agentes, abordagem técnica adotada neste trabalho. Primeiro, discute-se a fundamentação e a evolução da abordagem; em seguida, a lógica *bottom-up* responsável pelo comportamento emergente a partir de partes menores; e, por fim, as limitações da abordagem clássica.

2.2.1 Fundamentos de modelagem e simulação

De acordo com (Banks; Carson; Nelson; Nicol, 2010), simulação trata-se da imitação operacional de um processo ou sistema do mundo real ao longo do tempo. Com a simulação, é possível avaliar, através de diferentes entradas, as saídas resultantes e entender mais sobre o sistema sob investigação. O que diferencia a simulação da análise matemática estática é que a simulação é dinâmica. Com ela, é possível observar a evolução do estado de variáveis e suas interações nas diferentes etapas ao longo do tempo. A simulação é uma ferramenta que deve ser usada em análise de sistemas complexos, onde equações e análises estatísticas não são suficientes.

2.2.2 Evolução e Conceitos Fundamentais

Historicamente, a simulação de sistemas complexos evoluiu da análise estatística tradicional para a simulação computacional. Em ambiente comercial, modelavam-se apenas processos simples e lineares, como filas de caixa. Tais modelos não possibilitavam a captura da subjetividade das interações. Um dos ramos de modelagem e simulação de sistemas complexos é a modelagem e simulação baseadas em agentes (MSBA). Um agente é uma entidade computacional que é individual e autônoma. Sistemas multiagentes, por outro lado, consideram vários desses agentes interagindo em um ambiente compartilhado. Esses agentes são capazes de perceber, decidir e agir de forma independente no ambiente e entre si (Wooldridge, 2009).

2.2.3 Modelagem Bottom-up e Comportamento Emergente

Sua grande vantagem é que ela gera uma simulação complexa a partir de unidades constituintes menores. Através das microinterações dos agentes, é possível analisar comportamentos emergentes (Bonabeau, 2002), como, por exemplo, a dinâmica de consumo dentro de um ambiente de varejo. Essa abordagem bottom-up provou ser superior para capturar comportamentos individuais heterogêneos e autônomos, sendo hoje muito usada na modelagem de fluxos de multidões e disseminação de informações (Bonabeau, 2002). Segundo Siebers; Macal; Garnett; Buxton; Pidd, 2008, apesar de a simulação computacional ter sido utilizada fortemente desde os anos 1960, a MSBA teve seu uso expandido a partir da década de 1990, quando foi descrita não apenas como tecnologia, mas também como paradigma metodológico.

2.2.4 Limitações dos Agentes Clássicos

Contudo, os agentes tradicionais operam com base em regras simples, o que limita sua capacidade de reproduzir a complexidade do raciocínio humano, que é frequentemente imprevisível. Abordagens clássicas de SMA frequentemente empregam modelos de escolha discreta, que são abstratos e não incorporam traços de personalidade, preferências subjetivas e emoções ilógicas de forma satisfatória. Essa abstração gera agentes cujos comportamentos são difíceis de explicar ou validar em nível humano (Kato; Oshima; Nomura; Sugawara, 2024). Visto que um dos objetivos centrais da Inteligência Artificial é o desenvolvimento de sistemas que mimetizem o comportamento humano (Russell; Norvig, 2021), a integração de LLMs a esse tipo de modelagem dá origem a agentes capazes de ultrapassar barreiras comportamentais de arquiteturas anteriores.

2.3 Inteligência Artificial e Grandes Modelos de Linguagem (LLMs)

Esta seção explora a evolução da inteligência artificial, escopo em que o presente estudo se enquadra, da IA simbólica à IA moderna, vertente que possibilitou a criação dos grandes modelos de linguagem, motor da cognição generativa.

2.3.1 IA Simbólica

O termo Inteligência Artificial (IA) foi inicialmente proposto em 1956 por John McCarthy e posteriormente consolidado pelo mesmo como: a ciência e a engenharia de fabricar máquinas inteligentes com capacidade de simular características da cognição humana (Mccarthy, 2007). Entre o início dos anos 50 e as décadas posteriores, o paradigma da IA foi definido como Simbólica (Newell; Simon, 1976), uma vez que acreditava-se que a inteligência do comportamento humano poderia ser replicada atra-

vés da manipulação de símbolos e uso de regras lógicas. Nessa época, os programas eram totalmente dependentes da operação dos humanos para a inserção do conhecimento.

2.3.2 IA Conectivista

Durante essa mesma época, outra vertente do paradigma evoluía, porém, de forma mais lenta, ocasionada por fatores como complexidade, lacuna de ferramentas disponíveis, falta de dados e poder computacional exigido. Essa outra abordagem, chamada de IA conectivista (Rumelhart; Mcclelland, 1986), baseava-se na emulação de estruturas biológicas humanas, focando especialmente nos neurônios e sinapses do cérebro humano. O conhecimento não seria armazenado em apenas uma estrutura, mas sim em estruturas conectadas como pequenas unidades de processamento, os neurônios artificiais. Essa vertente, diferente da simbólica, não dependia da inserção manual de regras e símbolos, utilizando redes neurais artificiais para identificar padrões nos dados. Ainda assim, as redes dessa época ainda exigiam uma preparação manual dos atributos (do inglês, *feature engineering*); a extração de padrões diretamente dos dados brutos só veio com a IA moderna.

2.3.3 IA Moderna

Houve um grande avanço nessa área com o desenvolvimento da tecnologia e ascensão do aprendizado de máquina (do inglês - *Machine learning*) e do aprendizado profundo (do inglês - *Deep learning*) (Lecun; Bengio; Hinton, 2015), consolidando-se como a IA moderna. Essa abordagem não necessitava mais de programação explícita; ela conseguia extrair conhecimento de forma autônoma a partir de exemplos. Foi a vertente moderna que possibilitou a criação dos grandes modelos de linguagem, que são tecnologias que utilizam redes neurais complexas. Majoritariamente baseados na arquitetura *transformer* (Vaswani; Shazeer; Parmar; Uszkoreit; Jones; Gomez; Kaiser; Polosukhin, 2017), esses modelos conseguem captar nuances e contextos na linguagem humana, que seriam impossíveis de serem compreendidos utilizando mapeamento de símbolos tradicionais, provando-se uma abordagem relevante no apoio à tomada de decisão em ambientes complexos e dinâmicos (Russell; Norvig, 2021; Shankar, 2018). Ademais, a IA possui como objetivo central a geração de sistemas com capacidade de reprodução dos aspectos da inteligência humana, tanto em sua capacidade de raciocinar quanto no aprendizado e na interação com o ambiente (Russell; Norvig, 2021).

2.4 Agentes Generativos

Nesta seção, são definidos os agentes generativos, paradigma fundamental para a simulação proposta neste trabalho. O texto aborda o conceito desses agentes, detalha sua arquitetura cognitiva (composta pelos módulos de memória, reflexão e planejamento) e, por fim, explora as técnicas de engenharia de *prompt*.

2.4.1 Definição e Paradigma

O avanço da Inteligência Artificial possibilitou a criação de um novo paradigma de agentes. Agora, os agentes chamados generativos utilizam LLM para chegar a um nível superior de autonomia e raciocínio contextual (Xi et al., 2023). Aprimorando a SMA, esses agentes tentam superar parte da lacuna comportamental dos agentes clássicos. Esses agentes se diferenciam em três pilares principais: memória, reflexão e planejamento. Isso significa que eles conseguem armazenar experiências, refletir sobre experiências passadas e, através disso, planejar as ações futuras (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023 *apud* Kato; Oshima; Nomura; Sugawara, 2024). Isso deriva em um comportamento crível e coerente, que demonstra capacidade de aprendizado e adaptação, assemelhando-se à tomada de decisão humana, atingindo um nível de realismo comportamental antes considerado inalcançável (Kato; Oshima; Nomura; Sugawara, 2024; Wang; Xie; Jiang; Mandlekar; Xiao; Zhu; Fan; Anandkumar, 2023).

2.4.2 Arquitetura Cognitiva: Memória, Reflexão e Planejamento

Os dados que esses agentes armazenam funcionam simulando a memória humana, possuindo um histórico de diálogos, percepções e ações que o agente vive. Essa memória é finita e é constantemente atualizada com base na importância, relevância e atualidade, de acordo com o contexto (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023). A reflexão utiliza o módulo de memória; é responsável por revisitar esses dados brutos e gerar novas conclusões. Por fim, a partir dos dados antigos e dessas conclusões geradas, o agente consegue planejar seu próximo conjunto de ações (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023). Esse conjunto não é estático, podendo ser mudado de acordo com novas memórias e percepções.

Apesar desses avanços, os agentes generativos herdam limitações inerentes aos LLMs. A mais crítica é a alucinação, a geração de informações incoerentes ou factualmente incorretas apresentadas como verdadeiras, fenômeno observado inclusive em simulações de agentes, nas quais o acúmulo de memórias levou à criação de fatos que não ocorreram de fato (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023). Somam-se a isso o alto custo computacional e financeiro das inferências, a latência de resposta e o comportamento não determinístico.

2.4.3 Engenharia de Prompt

A engenharia de *prompt* (do inglês, *prompt engineering*) trata-se de projetar e formatar a entrada enviada ao modelo de linguagem, o *prompt*, a fim de guiar o comportamento e a qualidade da saída sem alterar os pesos do modelo (Liu; Yuan; Fu; Jiang; Hayashi; Neubig, 2023). Essa abordagem explora a capacidade de aprendizado dos LLMs com base no contexto fornecido na entrada (*in-context learning*), condicionando suas respostas com o uso de instruções e exemplos.

Há dois mecanismos centrais na engenharia de *prompt*. O primeiro, o *system prompt*, é uma instrução com maior prioridade, sempre adicionada ao contexto independentemente da entrada. Nela, geralmente se definem o papel, o estilo e as restrições que o modelo deve seguir ao longo da interação e, como no caso do presente trabalho, uma persona e um tom de resposta. O segundo é a injeção de contexto, na qual informações relevantes são adicionadas de forma dinâmica ao *prompt* para que o modelo consiga raciocinar sobre elas. Neste trabalho, esse contexto inclui, por exemplo, a situação atual da visita, os produtos na cesta, os setores por onde o cliente passou, os pensamentos sobre a conversa e o contexto da loja, entre outras informações.

A restrição das respostas é também um recurso importante para mitigar alucinações, saídas que possuem informações incoerentes ou factualmente incorretas. Ao ancorar o modelo em um contexto factual e restringir o formato esperado da resposta, reduz-se a tendência do LLM a inventar dados. Neste trabalho, a engenharia de *prompt* é o instrumento que dá origem às personas dos agentes cliente e vendedor, define o tom do atendimento e mantém as decisões ancoradas no catálogo real da loja, evitando a criação de produtos e preços inexistentes.

3 REFERENCIAL TECNOLÓGICO

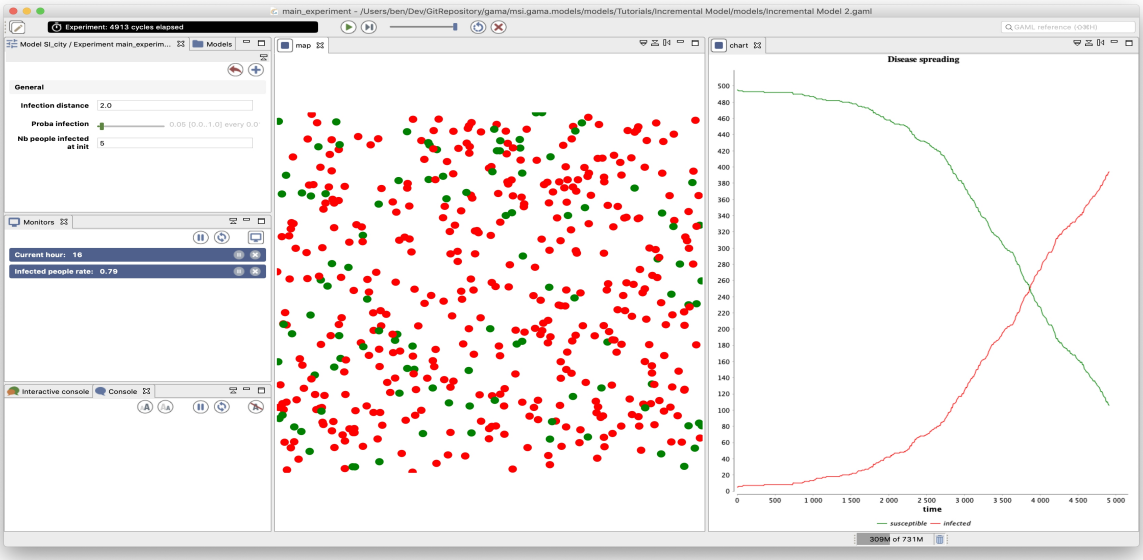
Este capítulo aborda uma breve explicação das ferramentas que foram utilizadas para o desenvolvimento do presente trabalho.

3.1 GAMA

GAMA é um software de simulação multiagente de código aberto que, conforme amplamente documentado por (Taillandier; Gaudou; Grignard; Huynh; Marilleau; Caillou; Philippon; Drogoul, 2019), permite modelar e analisar sistemas complexos por meio das interações entre os agentes. Essa ferramenta utiliza a linguagem GAML, que é declarativa, de alto nível e orientada a agentes. A estrutura dessa ferramenta divide-se em três partes fundamentais: *Global*, que dita o ambiente em que os agentes serão posicionados; *Species*, que representam os agentes e possuem atributos, ações e comportamentos ou reflexos executados de forma automática; por último, a seção *experiments* define como a simulação será executada, seja via interface gráfica (GUI) para visualização ou em modo *Batch* para processamento massivo de dados. Nesse projeto, ele será responsável pela simulação da abordagem clássica e também pelo motor espacial para a abordagem generativa, gerindo o *layout* da loja, mudanças no ambiente e as movimentações dos agentes.

A escolha do GAMA como motor espacial justifica-se devido a características relevantes para este trabalho. A plataforma é atualizada e possui utilização consolidada no meio científico, favorecendo a reprodutibilidade. O modo de execução sem interface gráfica (*headless*) permite grande processamento de dados, nesse trabalho, foi necessário para a calibração de parâmetros através da média de execução de diferentes seeds. Por fim, a ferramenta é flexível quanto à comunicação com frameworks externos, nesse caso de LLM, que era requisito central desse projeto. A Figura 1 apresenta a interface da plataforma.

Figura 1 – Visual da plataforma GAMA

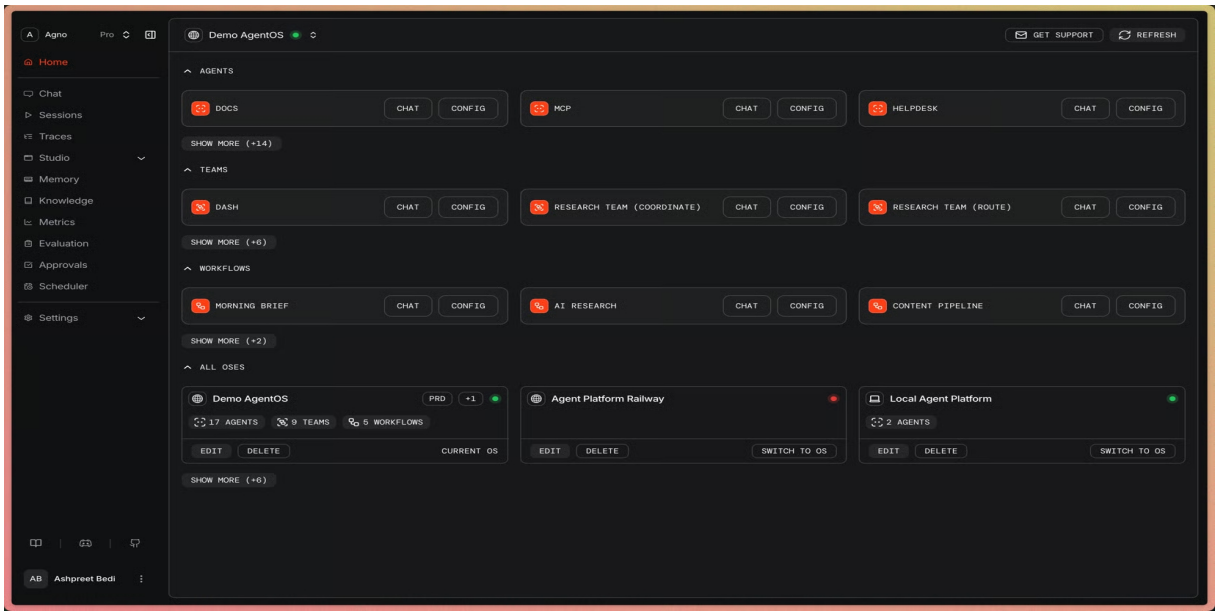


Fonte: `https://gama-platform.org/`

3.2 Agno

Agno é um framework de código aberto construído na linguagem Python, com ele é possível a construção de agentes inteligentes com módulos de memória e conhecimento. A ferramenta se destaca pela leveza, por depender de poucas bibliotecas e permitir a instanciação rápida de agentes, e pela modularidade, já que sua estrutura se divide em componentes independentes que podem ser combinados conforme a necessidade; essas características facilitam o uso de engenharia de *prompt* para definir as diferentes personalidades dos agentes. A estrutura do Agno se dá da seguinte maneira: SDK são as primitivas para criação de agentes e ferramentas que eles podem executar; *Engine* e AgentOS são camadas que gerenciam o ciclo de vida dos agentes; *Memory* ou memória é responsável pelo armazenamento persistente do histórico de interação e diálogo dos agentes, garantindo a continuidade do contexto do raciocínio em simulações mais longas, esses dados são armazenados em bancos de dados como SQLite ou PostgreSQL; *Knowledge* ou conhecimento é a base de dados necessária para que um agente tome uma decisão fundamentada em dados do domínio de aplicação(Agno, 2024). A Figura 2 ilustra a estrutura do *framework* (AgentOS).

Figura 2 – Visual do framework AGNO (AgentOS)



Fonte: `https://docs.agno.com/`.

3.3 API Agentic MAS

A API Agentic MAS é responsável pela integração das duas ferramentas utilizadas neste projeto. Desenvolvida como infraestrutura de propósito geral e descrita no artigo (Alves; Born; Aguiar, 2026), essa API realiza a comunicação entre os motores cognitivo e espacial através de requisições REST com mensagens em formato JSON. A cada momento que um agente ativa um gatilho e precisa raciocinar, o GAMA envia o contexto da situação, a API reencaminha para a LLM que responde com a decisão baseada no contexto e memória, a persistência das informações é garantida através de um banco de dados PostgreSQL.

A API realiza a organização da comunicação através de quatro entidades: *round* retém as métricas e contexto da rodada; *simulation* registra informações gerais da simulação como o número de agentes, tempo total de processamento, etc; *decision* é responsável pelo armazenamento das decisões tomadas pela LLM que serão reproduzidas pelos agentes; e *agent* que gerencia os atributos dos agentes simulados, e também armazena seu histórico permitindo uma análise por rodada. Como é uma arquitetura generalizada, neste projeto houve algumas adaptações para o caso de estudo que é o varejo, para viabilizar a lógica de clientes, vendedores, os tipos de decisão e formatos das respostas esperadas, descritas no capítulo de desenvolvimento.

4 TRABALHOS RELACIONADOS

Este capítulo apresenta, em uma revisão narrativa, os trabalhos relacionados a esta pesquisa, selecionados por conveniência e por similaridade com as tecnologias e o domínio deste estudo. Eles foram divididos em três grupos: agentes generativos, agentes generativos aplicados ao consumo e varejo, e dinâmica espacial da movimentação dos consumidores. Ao final há um resumo que trata do posicionamento do presente trabalho em comparação com os demais.

4.1 Agentes generativos

Os agentes generativos, descritos no referencial teórico, diferenciam-se dos agentes clássicos devido a sua capacidade cognitiva, que lhes permite planejar ações com base na reflexão do contexto, através do uso de LLM e módulos de memória. Os trabalhos compõem o conjunto estudado como base para o assunto, e ilustram diferentes aplicações desse paradigma.

4.1.1 Park et al. (2023) — Generative Agents: Interactive Simulacra of Human Behavior

O trabalho desenvolvido por (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023) teve como objetivo desenvolver uma cidade de agentes generativos e verificar se esses agentes, ao longo da simulação, conseguem armazenar o contexto de forma correta, assim como simular ações individuais e dinâmicas sociais mantendo a coerência a longo prazo. Apelidada de Smallville, a cidade conta com vários ambientes que uma cidade real possui: bares, cafeterias, casas, lojas, escolas e objetos com visual no estilo *pixel art*. A utilização de uma arquitetura com LLM permite a criação dos agentes, onde cada agente atua de maneira diferente de acordo com sua personalidade, definida por uma descrição em linguagem natural (*seed memory*) abrangendo aspectos como família, profissão, interesses e contexto. As suas memórias são armazenadas em linguagem natural e utilizadas durante o raciocínio dos agentes para geração de novas conclusões com base em prioridades de recência, importância e relevância de-

finidas pelos autores. O desempenho dos agentes como proxies humanos foi avaliado através de dois métodos: avaliação controlada, na qual os agentes foram entrevistados e suas respostas misturadas com respostas de seres humanos interpretando os personagens e com diferentes arquiteturas (com diferentes configurações de reflexão, planejamento e memória), sendo as respostas classificadas pelos avaliadores em naturalidade e coerência através de pontuação média e desvio padrão; e avaliação de ponta a ponta, em que a simulação foi rodada durante um longo período, sendo avaliada a credibilidade dos comportamentos emergentes não roteirizados. O resultado da avaliação controlada indicou que a arquitetura completa se destacou das outras abordagens, sendo responsável pela maior credibilidade das respostas. Já na avaliação de ponta a ponta, verificou-se que a difusão de informações ocorreu de forma orgânica, na qual eventos, como a candidatura do agente a prefeito, espalharam-se entre os agentes; além disso, a quantidade de relacionamentos entre agentes cresceu, o que foi demonstrado pelo aumento da densidade da rede de contatos na cidade. Apesar do sucesso nessas métricas, o estudo também encontrou limitações nessa abordagem. Verificou-se que o acúmulo de memórias causou comportamentos erráticos, como: agentes indo almoçar no bar, tentando entrar nas lojas que já estavam fechadas e outros comportamentos anormais, assim como alucinações em que os agentes embelezavam memórias e criavam fatos que não aconteceram realmente.

4.1.2 Kato et al. (2024) — Constructing LLM-based Agent Model and Its Application to Multi-Agent Simulation

Simulações multiagente clássicas possuem altos custos de desenvolvimento, especificamente na dificuldade de capturar e reproduzir a individualidade do comportamento humano de forma fiel e não enviesada. (Kato; Oshima; Nomura; Sugawara, 2024) possui como premissa a utilização de grandes modelos de linguagem a fim de abstrair a complexidade das simulações multiagente, acreditando que o LLM já é um grande repositório de conhecimento coletivo com dados do comportamento humano embutidos. Como a tomada de decisão é realizada em linguagem natural, é facilitada a análise da credibilidade do comportamento apresentado por esses agentes.

Para comprovação da premissa, foram realizados testes em um ambiente que simulava uma rua de restaurantes. Nessa rua, foram posicionados 1000 agentes que deveriam escolher em qual dos 11 restaurantes eles iriam almoçar. Os restaurantes servem comidas diferentes, como, por exemplo, italiana, japonesa, chinesa, etc., e possuem uma fila. Dentre os agentes, apenas 100 deles eram controlados por LLM, os outros utilizavam premissas e regras de escolha de restaurante aleatórias e simulação de espera nas filas. Os agentes generativos, por outro lado, deveriam escolher seu restaurante raciocinando com base em três fontes: perfil, informações do ambiente e memória. O perfil era informações pré-geradas que detalhavam a personalidade do

agente; ele continha nome, idade, traço inato de personalidade, aprendizado, comida favorita, estilo de vida e requisito do plano de almoço. As informações do ambiente eram dados dos cardápios dos restaurantes, o preço médio dos pratos e também o número de pessoas que estavam na fila. Por outro lado, a memória era registro em texto das experiências recentes do agente, lugares que havia olhado ou pensamentos que tinha tido antes, escritos em linguagem natural.

O estudo avaliou a credibilidade das decisões lógicas com base na personalidade, no raciocínio e na memória que eram apresentadas em linguagem natural. Foi identificado que o uso de linguagem natural tornou as ações altamente explicáveis e críveis comparadas com a tomada de decisão humana. Foi analisado que, ao adicionar diferentes características nos agentes, o efeito foi traduzido em ações coletivas reais na simulação, como, por exemplo, a diminuição de aglomeração nas filas ao adicionar percepções negativas a grandes filas nos traços de personalidade dos agentes. Além disso, foi comprovado que o resultado foi independente do modelo de IA utilizado, pois a mesma simulação utilizou diferentes modelos de LLM gerando resultados muito parecidos. O estudo mostrou que havia certa flexibilidade nos prompts sem alterar completamente os resultados, não sendo necessária uma engenharia de prompt complexa. A escalabilidade do sistema foi testada com diferentes porcentagens dos 1000 agentes sendo de LLM, e todas geraram resultados plausíveis; logicamente, quando todos os agentes foram de LLM, o tempo necessário para uma execução completa da simulação foi maior devido ao maior número de requisições. Em um teste isolado de interação entre esses agentes, foi analisada a possibilidade de eles chegarem a um consenso de forma autônoma na escolha do restaurante. Apesar de ser considerada uma tarefa social complexa, os agentes conseguiram concordar na sua escolha com algumas ressalvas: em conversas em que havia muitas discordâncias entre os agentes, geraram-se alucinações, como o chamado “restaurante de fusão”, um restaurante que não estava descrito nas regras, mas que, devido à tentativa de resolução de problema, foi inventado para comportar a necessidade de todos para o almoço.

4.2 Generativos aplicados ao consumo e ao varejo

Os trabalhos nessa seção aplicam agentes generativos no comportamento de consumo do varejo, tratam também de previsão de decisões de compra e avaliação de atendimento, se aproximando do objetivo de suporte à decisão pretendido nesse trabalho.

4.2.1 Mansour et al. (2025) — PAARS: Persona Aligned Agentic Retail Shoppers

O trabalho de (Mansour; Perelli; Mainetti; Davidson; D’amato, 2025) teve como objetivo avaliar se compradores de varejo agênticos alinhados à persona (PAARS —

do inglês *Persona Aligned Agentic Retail Shoppers*) conseguem simular o comportamento humano nas decisões de compra do ambiente virtual de *e-commerce*. O sistema cria esses agentes de compra automaticamente através de dados de compras reais. Primeiro, é analisado o histórico de compra e deduzido o perfil demográfico, englobando características como faixa etária, renda, estado civil, gênero, ocupação, etc. Então, com o perfil demográfico pronto, é realizada uma segunda análise mais profunda, responsável pela descrição de gatilhos psicológicos e comportamentais desse agente, determinando como o agente lida com sensibilidade a preço, dependência de avaliações, escolha de marcas e percepção de valor. Para a avaliação da credibilidade de mimetização do comportamento humano desses agentes, foi utilizada uma amostra de pessoas reais, e tanto os PAARS quanto os seres humanos foram submetidos a três tarefas do fluxo natural do *e-commerce*: geração de consultas, seleção de itens e geração de sessões. O desempenho foi medido em nível de grupo e individual. No quesito individual, a replicação de ações foi medida através de métricas de precisão e equivalência textual. Já no nível de grupo, as amostras foram avaliadas através da divergência de Kullback-Leibler, uma medida de divergência que quantifica a dissimilaridade estatística e a perda de informação pelo grupo de agentes em comparação com a amostra humana. Os testes foram feitos com diferentes abordagens de agente generativo, nas quais os agentes com persona se mostraram superiores aos modelos genéricos, evidenciando que houve redução do viés através dessa abordagem.

Após a calibragem e validação do modelo, o PAARS serviu como um *sandbox* auxiliando na previsão de resultados de testes A/B no meio digital, antecipando como os consumidores reagiriam a mudanças em preços, disposição de elementos no site ou algoritmos, sem a necessidade de implementação na realidade.

4.2.2 Li, Wei & Wang (2025) — Predicting Behaviors with Large Language Model (LLM)-Powered

O trabalho de (Li; Wei; Wang, 2025) teve como objetivo a criação de gêmeos digitais (entidades virtuais que procuram replicar alguma entidade do mundo real); neste caso, o foco era mimetizar o comportamento do consumidor. Para isso, foram usados dados demográficos, comportamentais, psicográficos e transacionais de perfis em sites de compra. O processamento de dados e criação dos agentes generativos fizeram uso de tecnologias como LLMs e Geração Aumentada por Recuperação (RAG, uma técnica que melhora a precisão e reduz alucinações dos LLMs). A avaliação da credibilidade do sistema se deu através da utilização de perfis com no mínimo 400 avaliações de compras na plataforma Amazon Shopping. Foram utilizados 304 perfis de usuários e, para cada perfil, foi criado o seu gêmeo digital. As compras e avaliações foram utilizadas para treinamento do modelo, com exceção de uma amostra das últimas 50 compras de cada usuário. Essas 50 compras foram embaralhadas com

outros 50 produtos que não haviam sido adquiridos por ele. O gêmeo digital deveria prever, através da sua personalidade e informações, se o usuário compraria aquele item e, em caso positivo, realizar uma avaliação do produto. O sistema gerou cerca de 86% de acerto na previsão de compras dos clientes verdadeiros e, no teste em que os agentes produziam avaliações de produtos, alcançou-se cerca de 0,94 de similaridade de cosseno, onde 1 significa a correspondência semântica e lexical perfeita com as avaliações reais dos usuários.

4.2.3 Oliveira (2025) — Desenvolvimento de um sistema multiagente baseado em inteligências artificiais generativas para avaliação de qualidade em atendimento ao cliente

O trabalho de conclusão de curso realizado por (Oliveira, 2025) é, possivelmente, a abordagem nacional que mais se aproxima do presente estudo. Oliveira construiu um sistema multiagente utilizando LLMs, cujo objetivo foi empregar agentes generativos para avaliar a qualidade dos atendimentos em centrais de relacionamento com o cliente. Dado que a prestação de serviço está ligada diretamente ao usuário final, o cenário empregado se encaixa na definição de varejo. O sistema era composto por uma equipe de cinco agentes: Supervisor, Monitor de Qualidade, Operador, Juiz e Analista de Relatórios, cada um responsável por aspectos específicos da avaliação. Para os testes, foram utilizadas conversas reais de uma empresa de energia, o modelo de linguagem GPT-4o-mini e o *framework* CrewAI. A solução proposta pelo autor obteve resultados expressivos de precisão em comparação ao uso de uma IA isolada, além de demonstrar ganhos em eficiência de tempo, redução de custos e qualidade em relação à avaliação humana. No entanto, o autor apontou algumas limitações em sua abordagem: a necessidade de a transcrição já ser fornecida pronta (mantendo a dependência de um processo de conversão prévio) e a ausência de processamento de áudio, o que impede a captação do tom de voz e de outros maneirismos humanos cruciais em determinadas situações.

4.3 Dinâmica espacial e movimentação de consumidores

Os trabalhos a seguir tratam da dinâmica do ambiente físico do varejo, isto é, movimentação e navegação de clientes, assim como análises de densidades de calor por departamentos.

4.3.1 Horikomi & Mizuno (2025) — Interaction-aware agent-based simulation of customer trajectories in retail stores with transformer architectures

Para simular a movimentação de consumidores em ambiente de varejo físico com alta fidelidade, o trabalho de (Horikomi; Mizuno, 2025) propôs um sistema que utiliza:

rastreamento, técnicas de conversão de coordenadas físicas em texto e a arquitetura *Transformer*, tecnologia que possui a mesma base de funcionamento dos LLMs, mas que se difere por não processar a linguagem natural humana. Esse tipo de arquitetura deve ser treinada do zero para processar uma linguagem; no caso do autor, a linguagem foram as coordenadas transformadas em texto, permitindo a previsão das trajetórias, desvios, desaceleração, pausas e desengajamento precoce devido a interações entre agentes. O sistema foi avaliado através da comparação entre trajetórias de consumidores previstas pelo modelo e trajetórias de humanos reais coletadas por sistemas de rastreamento localizados em estabelecimentos. Os autores fizeram uma bateria de testes com 8 diferentes escalas de aglomeração, variando de situações de loja vazia até um fluxo de 720 clientes por hora. A fidelidade, então, foi medida através de 6 métricas: distribuição de densidade espacial, duração de permanência, distância percorrida, ângulo de giro, velocidade vs. densidade local e erro quadrático médio. Para provar que a movimentação do modelo era dependente da interação entre agentes, foi alterada a quantidade de 'vizinhos' considerados na simulação. Os resultados demonstraram que o modelo conseguiu replicar com precisão os padrões reais em todas as seis métricas avaliadas. Além disso, constatou-se que a simulação alcançou sua maior fidelidade (menor margem de erro) ao considerar as ações de exatamente dois clientes mais próximos. Indicando possível viés de treinamento, já que as situações de aglomeração com 4 pessoas ou mais eram raras, fazendo com que o modelo não soubesse como se comportar da melhor forma quando havia esse vizinho a mais. Independentemente, o estudo comprovou empiricamente que a análise do contexto social local é indispensável para gerar movimentos críveis no espaço físico.

4.3.2 Pantano et al. (2021) — Enhancing store layout decision with agent-based simulations of consumers' density

(Pantano; Pizzi; Bilotta; Pantano, 2021) realizaram um estudo sobre como o layout afeta a densidade de consumidores nas zonas de um estabelecimento. Para isso, foi utilizada uma simulação multiagente com mecanismo de Monte Carlo, calibrada através de dados de um supermercado localizado dentro de um shopping na Itália. Os dados permitiram identificar o fluxo de clientes em diferentes horários e dividir os setores desse mercado com base em sua atratividade. Foram simulados dois cenários no software Wolfram: o primeiro, sendo o cenário real que acontecia diariamente, e o segundo com o layout reorganizado de modo que as zonas com maior atratividade ficassem mais distantes entre si. A eficácia dessa reorganização foi avaliada através da função de densidade de probabilidade em conjunto com a entropia (grau de desordem), mostrando que a reorganização de layout diminuiu gargalos e picos de aglomeração, e distribuiu os clientes de maneira equilibrada pelo espaço. O estudo apresentou algumas limitações: o número de clientes foi baseado nos recibos, ou seja,

não considera os clientes que não compraram; além disso, a simulação se manteve no âmbito digital, não houve mudança verdadeira de layout no espaço físico, logo não há dados quanto à percepção de mudança de layout por consumidores.

4.4 Posicionamento do Trabalho Proposto

O presente trabalho herda elementos dos trabalhos relacionados, diferindo-se pela união de seis eixos que, apesar de serem tratados pela literatura, não aparecem reunidos nos trabalhos analisados: varejo, simulação multiagente, cognição de agentes com LLM, predição, validação com dados reais e utilização de espaço físico. No âmbito do uso de LLMs em simulação multiagente, (Kato; Oshima; Nomura; Sugawara, 2024) e (Park; O'brien; Cai; Morris; Liang; Bernstein, 2023) o fazem em ambientes sociais genéricos, sem o elemento varejo incluso. Já quanto aos trabalhos onde há ambos os elementos, (Mansour; Perelli; Mainetti; Davidson; D'amato, 2025) e (Li; Wei; Wang, 2025) realizam análise exclusiva do comportamento do consumidor no varejo digital, enquanto (Oliveira, 2025) utiliza a tecnologia para fazer uma avaliação de chamadas com o cliente final.

Este projeto, por outro lado, simula os consumidores em uma loja com dimensionamento baseado em uma loja de vestuário real, incorporando a complexidade das interações sociais relacionadas com o ambiente físico. Quanto ao estudo do espaço físico no varejo, o trabalho de (Horikomi; Mizuno, 2025) trata da simulação fiel da trajetória complexa em ambiente de varejo físico, utilizando a tecnologia *Transformer*, que possui a mesma base dos LLMs, enquanto (Pantano; Pizzi; Bilotta; Pantano, 2021) analisam a densidade de pessoas em uma loja e prediz se alterações no *layout* melhorarão a passagem de pessoas. A Tabela 1 sumariza e compara os trabalhos estudados como referencial para este estudo.

O presente trabalho contará com movimentação de agentes, mas como um instrumento e não como objeto de estudo; a movimentação dos agentes estará ligada à cognição dos agentes e aos perfis de consumidor. A presença desse ambiente físico permite a análise de indicadores espaciais exclusivos do varejo tradicional, descritos por (Underhill, 2009), como: *dwell time*, densidade de permanência por departamento e análise de gargalos na formação de filas. Conceitos correlatos como o *butt-brush effect* e zona de descompressão contextualizam a modelagem ainda que não medidos.

Tabela 1 – Comparativo entre os trabalhos.

Trabalho	**Foco no varejo**	**Usa SMA**	**Usa LLM**	**Espaço físico**	**Preditivo / contrafactual**	**Valida com dados reais**
Park et al. (2023)	Não	Sim	Sim	Não	Não	Não
Kato et al. (2024)	Não	Sim	Sim	Não	Não	Não
Mansour et al. (2025)	Sim	Não	Sim	Não	Sim	Sim
Li, Wei & Wang (2025)	Sim	Não	Sim	Não	Sim	Sim
Oliveira (2025)	Sim	Sim	Sim	Não	Não	Sim
Horikomi & Mizuno (2025)	Sim	Sim	Não	Sim	Sim	Sim
Pantano et al. (2021)	Sim	Sim	Não	Sim	Sim	Sim
Este trabalho	**Sim**	**Sim**	**Sim**	**Sim**	**Sim**	**Sim**

Fonte: Elaborada pelo autor.

5 METODOLOGIA

Este capítulo descreve a metodologia adotada para a construção e a avaliação da simulação, apresentando uma visão geral da abordagem com os motores espacial e cognitivo, o estudo de caso, o processo de modelagem com fontes reais, a calibração e a validação do modelo e o desenho experimental que compara as duas abordagens, clássica e generativa.

5.1 Visão geral

Para responder à necessidade de testar técnicas e estratégias sem o risco operacional real, a metodologia desse projeto consistiu na modelagem de uma simulação multiagente, calibrada com dados reais de uma loja do ramo da moda localizada no Rio Grande do Sul. Foi avaliado o poder de representação do comportamento do consumidor com o uso de agentes clássicos baseados em heurísticas e regras, e também agentes de LLM que apresentam o uso de linguagem natural e conseguem raciocinar antes de realizar uma ação através de contexto com base em memória.

Para o motor espacial foi escolhida a ferramenta GAMA (Taillandier; Gaudou; Grignard; Huynh; Marilleau; Caillou; Philippon; Drogoul, 2019), devido à sua flexibilidade na ponte com frameworks de agentes generativos, atualização e uso em trabalhos científicos, enquanto para o motor cognitivo dos agentes generativos foi escolhido o framework Agno (Agno, 2024), devido à sua grande utilização para SMA e baixa curva de aprendizado.

5.2 Estudo de caso

O estudo de caso utiliza dados datados em abril de 2026, provenientes de uma loja de vestuário popular localizada no Rio Grande do Sul que, por questões de sigilo comercial, optou por não ter sua identidade divulgada neste trabalho. O ambiente da loja possui 961 m² e foi modelado no GAMA respeitando a sua planta. As zonas: infantil, tênis infantil, masculino, tênis masculino, feminino, tênis feminino, cameba (cama, mesa e banho) e beleza também foram posicionadas em suas respectivas localiza-

ções, com um erro médio de aproximadamente 9% em comparação com suas áreas reais. Os produtos foram modelados com base no Pareto, um grupo de 209 famílias de produtos que representa 90% das vendas, enquanto o comportamento/perfil dos agentes foi modelado através da união de fontes: percepção do gerente da loja, dados disponibilizados e literatura científica.

O mês de abril, que se manteve apenas 1,2º C acima da média anual de temperatura, foi escolhido devido à sua generalidade, uma vez que a sazonalidade é comprovadamente responsável pela variação nas vendas de lojas desse ramo. Dados reais desse mês, como por exemplo: conversão de 31,77%, *ticket* médio de R$ 229,12, item/cesta de 2,59 e a média de 998 entrantes por dia útil, foram tomados como alvos a serem atingidos através da modelagem sem enviesamento ou *overfitting*. A Tabela 2 sintetiza as fontes de dados reais utilizadas e o que cada uma calibra no modelo.

Tabela 2 – Fontes de dados que abastecem o modelo.

Fonte de dado	**O que fornece ao modelo**
Planta da loja	Geometria, zonas e áreas
Pareto de produtos	Catálogo de produtos vendidos
Faixas horárias	Curva de chegada por faixa de horário
Indicadores gerais	Âncora de conversão e entrantes; validação de ticket, itens/cupom e mix por setor
Levantamento do gestor	Equipe de vendedores e turnos; perfil do cliente em dias úteis
Literatura	Parâmetros sem dado local (dwell, velocidade, etc.)

Fonte: Elaborada pelo autor.

5.3 Modelagem inicial

A simulação teve como pilar garantir a honestidade de todo número e todo comportamento, assegurando que os dados fossem ancorados em fontes reais ou declarados honestamente como premissas, com objetivo de adicionar o menor número de premissas necessárias. As informações foram divididas em três tipos de evidência: fonte direta, com dados reais; premissa triangulada, em que o comportamento aparece em pesquisas similares, mas sem um valor específico para a calibração do modelo; e premissa do autor, quando se sabe que o comportamento existe, mas não está documentado, buscando sempre a fonte com maior credibilidade disponível.

Para evitar o *overfitting*, ao tratar de premissas, procurou-se fazer apenas calibrações legítimas e justificáveis. Nenhum parâmetro artificial desonesto foi incluído no modelo a fim de atingir os indicadores-alvo; a lacuna entre indicadores emergentes e de referência foi tratada como limitação do modelo. As premissas sem fonte emba-

sada em literatura foram submetidas a uma análise de sensibilidade um fator por vez (OAT): cada premissa variou isoladamente entre os extremos de uma faixa de valores estabelecida, e cada nível foi executado no modo de dia inteiro, com a média de dez *seeds*, para verificar a dependência da simulação em relação a esses parâmetros. A Tabela 3 apresenta os resultados dessa análise.

Tabela 3 – Análise de sensibilidade das premissas

Premissa	**Faixa**	**Conversão**	**Ticket (R$)**	**Itens/cupom**
Fração de visitantes (*browsers*)	0,50–0,80	50,1→19,5%	190→186	2,53→2,52
Meses entre compras de vestuário	1,5–3,5	29,5→33,0%	128→255	2,03→3,04
Probabilidade de compra (clássico)	0,40–0,65	32,8→32,9%	186→196	2,45→2,61
Tempo de avaliação por arara	40–120 s	34,1→32,9%	196→170	2,55→2,15
Tempo de permanência	30–42 min	32,9→33,0%	186→194	2,45→2,59
Fração do público feminino	0,51–0,69	32,8→32,9%	192→193	2,55→2,56

Fonte: Elaborada pelo autor.

Os intervalos de variação de cada parâmetro foram definidos segundo três critérios. Para parâmetros com fonte, adotou-se a faixa que a literatura sustenta. Para os botões de calibração, adotou-se uma banda simétrica em torno do valor calibrado. Para as premissas sem fonte fina, adotou-se uma banda ampla em torno do valor assumido, larga o suficiente para revelar eventual sensibilidade. A justificativa de cada intervalo encontra-se na Tabela 4.

Tabela 4 – Origem das faixas de variação de cada parâmetro

Parâmetro	**Faixa**	**Origem da faixa**
`frac_browser`	0,50–0,80	benchmark de loja física de moda (Hoxton AI / Int. Retail Academy); tipologia browser×buyer (Moe, 2003)
`frac_feminino`	0,51–0,69	piso censo IBGE (51%) a IEMI (69%, consumidoras de vestuário)
`orc_meses_compra`	1,5–3,5	banda em torno da mediana (2,5 meses) da distribuição (Opinion Box)
`p_compra_classico`	0,40–0,65	premissa mecânica (máxima entropia)
`permanencia_media`	30–42 min	dwell híbrido loja-de-departamento × *fast-fashion* (Placer.ai, 2024)
`dur_avaliar`	40–120 s	premissa mecânica (máxima entropia)

Fonte: Elaborada pelo autor.

A conversão é ancorada pelo eixo *buyer*/*browser*, onde o parâmetro fração-*browser*

dita quantas das pessoas que entram na loja são apenas visitantes/*browsers*, que vieram somente para olhar, cravando o teto de conversão, seu valor tem como fonte benchmarks de conversão de lojas físicas de moda (Hoxton AI e *International Retail Academy*); o *ticket* é sensível a dois parâmetros: o tempo de permanência do cliente na loja e o seu orçamento. O tempo de permanência (Underhill, 2009) não é medido na loja real, mas foi estimado através de pesquisas comerciais e da visão do gerente; já o orçamento foi estimado através da renda multiplicada pelo percentual de gasto em vestuário e pelo número de meses entre idas à loja, com a renda por décimos vinda da Síntese de Indicadores Sociais do IBGE (IBGE, 2023), o percentual de vestuário da Pesquisa de Orçamentos Familiares (Lima; Araújo; Silva, 2024) e a mediana de meses da Opinion Box (Opinion Box, 2024). A quantidade de itens por cesta, assim como o *ticket*, possui correlação positiva com o tempo de permanência e o orçamento. A simulação mostrou-se insensível à maior parte dos demais fatores, uma vez que variá-los dentro dos intervalos definidos alterou os KPIs menos do que o próprio ruído entre as sementes do *baseline*.

5.4 Calibração e validação

O processo de calibração do modelo utilizou a média da execução de diferentes *seeds* no modo *headless* de execução do GAMA, o que torna mais eficiente para grandes volumes de dados por não apresentar interface gráfica. Uma execução é equivalente a um dia inteiro de loja, das 08:30 às 19:00. Como cada ciclo da simulação corresponde a um segundo, os clientes podem entrar na loja até cerca de 38.000 ciclos (aproximadamente 10 horas e 30 minutos, o horário de funcionamento); os ciclos posteriores seguem até a loja esvaziar. Foi realizada uma análise anti-viés no modelo, onde foram identificados parâmetros artificiais adicionados somente para atingir a meta de indicadores, que foram posteriormente trocados por parâmetros ancorados em fontes reais.

Cada comportamento, antes de ser incluído no modelo, foi avaliado por meio de testes A/B com diferentes *seeds*, testando o seu impacto nos indicadores e avaliando o comportamento emergente para verificar se sua inclusão era justificável. O modelo final foi considerado validado após reproduzir de forma satisfatória os KPIs-alvo, dentro de uma margem de erro aceitável e conter apenas premissas justificáveis. Os indicadores que não atingiram o alvo foram documentados como um achado, evidenciando uma limitação, pois forçá-los por meio de um ajuste não ancorado ocasionaria *overfitting*.

5.5 Desenho experimental e métricas

A avaliação da simulação está relacionada a dois tipos distintos de métricas. O primeiro grupo de indicadores é responsável pela validação do modelo e reúne a taxa de conversão, o ticket médio, o número de itens por cupom, o mix de venda por departamento e o volume diário de entrantes, métricas que são confrontadas com os valores reais da loja. Apesar de a taxa de conversão e o volume diário de entrantes serem ancorados e não emergentes, eles indicam a saúde da simulação. O segundo grupo reúne métricas de diagnóstico, como o tempo de permanência por zona, a taxa de interceptação do vendedor e a formação de fila, comportamentos que emergem dos agentes e não possuem medição direta na loja real, mas que ajudam a aferir a plausibilidade do comportamento emergente e a comparar cenários. É equivocado tratar o segundo grupo como se fosse validado, pois significaria atribuir precisão a dados que não possuem base empírica comprovada.

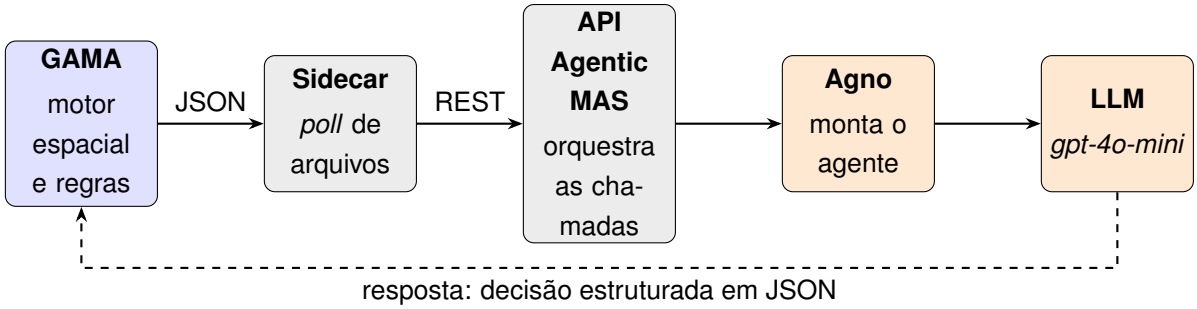
O experimento principal é a comparação entre as abordagens de agentes (clássicos e generativos). Em ambos os casos, a loja segue a mesma configuração: mesmo ambiente, dados e métricas. No grupo de controle, o comportamento dos agentes cliente e vendedor é pré-determinado, baseado em premissas fixas e regras probabilísticas. Já no grupo de tratamento, os agentes generativos utilizam de memória, raciocínio e persona (conjunto de características sorteadas estatisticamente), a LLM usa esses três pilares para decidir o motivo da ida à loja, para qual departamento os clientes navegam, quais itens compram, e também a conversa entre vendedor e cliente. O objetivo é executar a simulação nessas duas abordagens e comparar os indicadores emergentes com os indicadores alvo.

O segundo experimento é verificar a plausibilidade dos resultados da abordagem generativa, ao lidar com cenários em que há diferentes estratégias de gerenciamento em vigor, como mudanças no layout, inclusão de promoções e tipos de atendimento, permitindo estimar o efeito nos indicadores sem o real risco operacional. É importante ressaltar que, em ambos os experimentos, realiza-se a média de múltiplas *seeds* para reduzir o ruído estatístico entre execuções.

6 DESENVOLVIMENTO

Este capítulo detalha a modelagem e a implementação computacional da simulação. Para facilitar a compreensão da arquitetura do sistema, o texto foi estruturado da seguinte maneira: a Seção 6.1 descreve a configuração espacial, abordando o ambiente, o layout e as zonas. A Seção 6.2 explica as mecânicas de movimentação e navegação dos agentes pelo espaço físico, seguida pela Seção 6.3, que apresenta a modelagem do catálogo de produtos da loja. A Seção 6.4 detalha o controle de chegada e volume de fluxo dos clientes. Posteriormente, as características e as regras de decisão dos agentes são apresentadas nas Seções 6.5 e 6.6. Por fim, a Seção 6.7 explica a dinâmica das filas e do atendimento, e a Seção 6.8 apresenta a camada cognitiva com uso de LLM. A Figura 3 apresenta uma visão geral da arquitetura do sistema.

Figura 3 – Arquitetura do projeto: Motores espacial e cognitivo; e ponte.



Fonte: Elaborada pelo autor.

6.1 Ambiente, layout e zonas

A modelagem do ambiente físico começa com a utilização da planta arquitetônica real da loja (Figura 4), de onde foram extraídas a posição e dimensão das áreas reais da loja, que incluem: salão, caixa, departamentos e provadores. Para obter as medidas e coordenadas que não constavam no documento, foi utilizada a ferramenta MeasureAll, uma ferramenta de medição manual passível de erro, disponível em `https://measureall.app/en/`.

Figura 4 – Planta arquitetônica



Fonte: Disponibilizado pela loja - Feito por Arq. Marcelo Dalle Ore Braga.

A modelagem no GAMA seguiu a escala 1:1 (1 unidade de distância GAMA = 1 metro), com construção utilizando o formato retangular para formas simples e poligonal para representar áreas com contornos irregulares. Os seis departamentos modelados são Feminino, Masculino, Infantil, Calçados, Cameba e Beleza. A área do salão possui erro de menos de 1% na representação, enquanto os departamentos formados por múltiplas zonas e formatos complexos atingiram erros maiores como apresentado na Tabela 5, que não comprometem o comportamento da simulação.

Tabela 5 – Áreas das seções: modelo versus loja real.

Seção	**Modelo (m^2)**	**Real (m^2)**	**Diferença (%)**
Feminino (Superior + Inferior + Beleza)	243,5	267	−8,8
Masculino	133,7	160	−16,4
Infantil	136,9	150	−8,7
Calçados (feminino, masculino e infantil)	126,9	136	−6,7
Cameba	53,3	57	−6,5
Salão (contorno total)	969,4	961	+0,9

Fonte: Elaborada pelo autor.

O comportamento da zona de descompressão (Underhill, 2009), que indica que os produtos nos primeiros metros da entrada não são vistos pois o cliente está se acostumando com o ambiente foi modelado, mas não como uma zona, para isso utilizou-se uma premissa de tempo quando os clientes chegam à loja, eles não interagem com os produtos pois estão se adaptando. Os provadores reais da loja estão presentes,

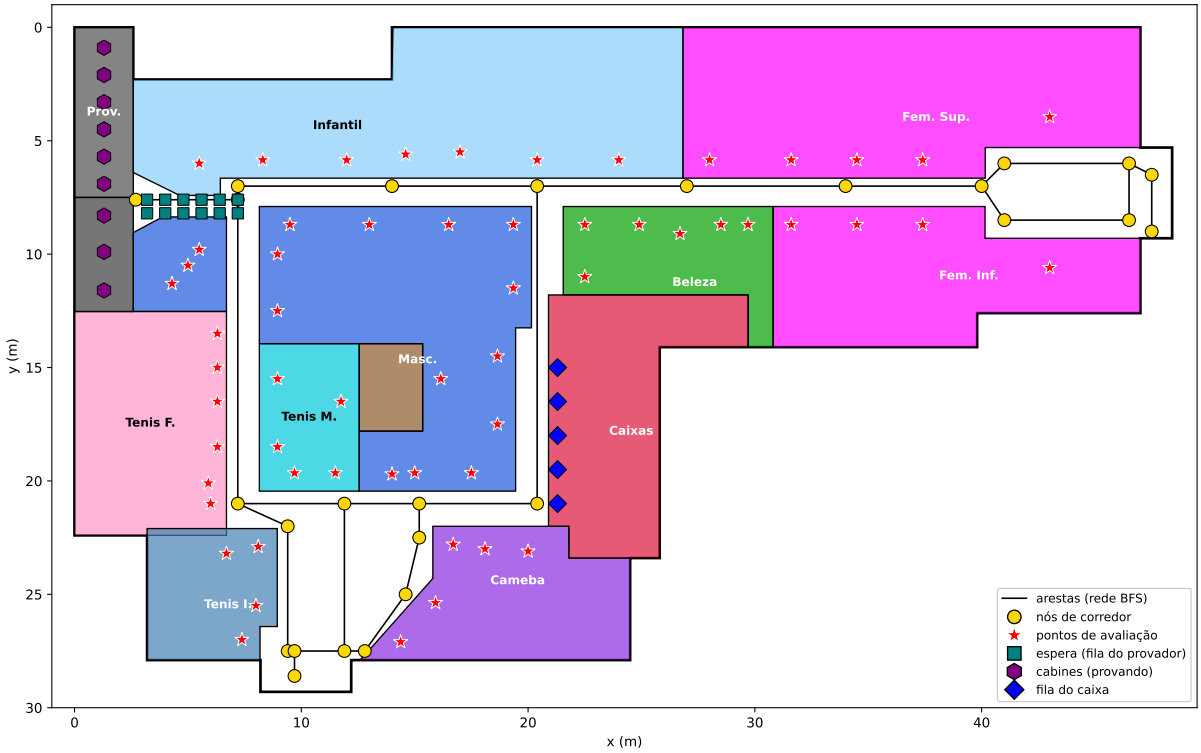
6 femininos e 3 masculinos, quando um cliente adiciona um item do catálogo que é da classe “prova_cabine”, em determinado momento da sua visita ele vai para o provador. A dinâmica dos caixas também foi modelada, os clientes que possuem itens na cesta vão ao caixa após o seu tempo de loja ter finalizado, ambos provadores e caixas possuem tempo de uso com premissas, e seus usos podem gerar filas devido ao número limitado de cabines e atendentes.

6.2 Movimentação e navegação

Para a movimentação e navegação dos agentes pelo ambiente, primeiro determinou-se a velocidade de deslocamento. O estudo de (Mohamad ali; Abustan; Abu talib, 2019), realizado em dois shopping centers, mediu a velocidade de caminhada de pedestres em ambiente comercial: embora em via pública os pedestres caminhem a aproximadamente 1,0–1,4 m/s, em ambiente de loja eles reduzem o passo. Além disso, observou-se que a velocidade média difere entre os sexos. Dessa forma, a velocidade dos agentes na simulação baseou-se nessa pesquisa, sendo 0,93 m/s para os homens e 0,81 m/s para as mulheres.

Para a navegação dentro da loja, optou-se por construir um grafo: o conjunto de nós foi posicionado seguindo a estrutura dos corredores e as arestas estabeleceram as conexões entre eles. Observou-se que, ao utilizar o método de movimentação nativo do GAMA (`goto on: grafo`), alguns agentes ficavam presos, oscilando junto às bordas das zonas sem alcançar o destino; a solução foi implementar um método próprio de busca em largura (BFS), o `caminho_ate`. O grafo resultante reúne 26 nós e 29 arestas. Quando um cliente sorteia um destino, no modo clássico, ou a LLM o decide, no modo generativo, o agente verifica se está próximo (a 6 m): caso esteja, anda em linha reta até o destino; caso contrário, recorre ao uso de corredores representados pelo grafo, realizando uma busca em largura para definir o menor caminho entre o nó de origem do cliente e o nó mais próximo do destino final. Quando o destino é um departamento, o destino do cliente é um ponto de avaliação; quando é no provador, um dos pontos de espera com o princípio FIFO (*First In, First Out*), até que uma das cabines do seu sexo esteja livre; e quando é no caixa, um dos pontos de fila, que também funcionam no estilo FIFO, até que um operador esteja disponível. A Figura 5 ilustra a rede de navegação sobreposta à planta da loja: os nós dos corredores (em amarelo), conectados pelas arestas (em preto) que formam o grafo de busca; os pontos de avaliação de cada departamento (estrelas vermelhas), os pontos de fila no caixa (losangos azuis), as 9 cabines de provador (hexágonos roxos) e os pontos de espera no provador (quadrados verdes).

Figura 5 – Grafo, arestas e pontos de avaliação



Fonte: Elaborada pelo autor.

6.3 Catálogo de produtos

A loja real possui um catálogo com milhares de produtos, com variações de tamanho, marca e cor, cada uma dessas variantes com seu código SKU (*Stock Keeping Unit*); para abstrair complexidade, foram simuladas famílias ao invés de cada SKU. Além disso, as famílias simuladas constavam no Pareto de abril, isto é, o grupo de itens responsáveis por aproximadamente 90% das vendas daquele mês, resultando em 209 famílias de itens modelados. Alguns exemplos de famílias desse grupo são: calça jeans adulto feminina, calça jeans adulto masculina, kit cueca adulto, tênis feminino adulto, pijama adulto feminino, etc. Cada família de item possui: código, nome e preço médio. A distribuição de famílias entre os 6 departamentos (Tabela 6) foi feita através dos prefixos do código e, para produtos ambíguos, foi decidida através de palavras-chave no nome, em que constavam: menino, menina, feminino, masculino, bebê, infantil. Já sua classe de avaliação é premissa.

Tabela 6 – Departamento de cada família pelo prefixo do código

Prefixo	Conteúdo	Área(s) de destino	Famílias
101	Roupa feminina	Feminino	42
102	Roupa masculina	Masculino	26
103 / 104 / 105	Roupa infantil	Infantil	62
106	Calçados	Feminino / Masculino / Infantil	32
107	Cameba	Cameba	17
108	Íntimo	Feminino / Masculino	13
109 / 112	Beleza / acessórios	Beleza	8
110	Pijama	Feminino / Masculino	9

Fonte: Elaborada pelo autor.

O estoque dos itens não foi modelado. Apesar de ter acesso aos dados, os números de estoque são atribuídos a cada variante de um item (cor, tamanho, etc.); logo, a abstração dos SKUs impossibilita atribuir uma magnitude geral de estoque para uma família. A classe de avaliação define a jornada de prova do cliente: itens sem prova não obrigam os clientes a irem ao provador; calçados são experimentados no próprio departamento; já roupas experimentáveis adicionam a necessidade de o cliente ir até o provador. Dependendo da jornada, o cliente vai ter mais ou menos tempo para avaliar outros itens, uma vez que seu tempo de loja vai passando. Quando um cliente vai a determinado departamento, o produto que o cliente avalia é aleatório dentre os que cabem no seu orçamento.

6.4 Chegada e volume

Para que a chegada de clientes se aproximasse do fluxo real da loja, utilizou-se o dado de entrantes por faixa de horário fornecido pela loja (Tabela 7). Considerando apenas os dias úteis (segunda a sexta), são 19.968 entrantes no mês que, divididos pelos 20 dias efetivamente abertos de abril (descontando os dois feriados em que a loja não operou), resultam em aproximadamente 998 clientes por dia útil. Esse valor foi tomado como alvo diário na simulação. O sábado não foi incluído devido a ser outro perfil de consumidor e ter volume bem diferente. Neste trabalho, modelou-se o dia útil típico.

Tabela 7 – Entrantes por faixa de horário e dia da semana (abril/2026)

Faixa horária	Seg	Ter	Qua	Qui	Sex	Sáb
08h às 09h	163	88	226	191	143	160
09h às 10h	362	340	570	476	276	900
10h às 11h	325	285	541	526	252	826
11h às 12h	259	182	339	388	116	569
12h às 13h	263	188	371	397	153	348
13h às 14h	322	289	490	568	309	426
14h às 15h	532	355	717	678	366	592
15h às 16h	527	340	642	673	358	644
16h às 17h	395	287	554	584	304	661
17h às 18h	332	257	585	508	348	377
18h às 19h	236	203	270	355	162	0
19h às 20h	0	1	0	1	0	–
Total	**3.716**	**2.815**	**5.305**	**5.345**	**2.787**	**5.503**

Fonte: Contador de fluxo da loja (abril/2026).

Para reproduzir os diferentes fluxos de pessoas nos horários, dividiu-se o tempo de simulação (08:30 – 19:00) em 11 janelas, sendo a primeira de meia hora (abertura) e as demais de 1h (Tabela 8). As janelas são ponderadas através da média de entrantes de cada horário, em dias úteis. Para calibrar o fluxo entre os horários, muda-se o intervalo médio entre as chegadas dos clientes. No final do dia, a quantidade de clientes deve resultar em aproximadamente 998 entrantes. Os valores de entrantes e peso do modelo que constam na tabela, foram medidos através da média de 10 seeds.

Tabela 8 – Janelas de chegada: intervalo, volume e peso de cada janela no total do dia

Janela	Duração	Intervalo	Entrantes	Peso real	Peso modelo
08:30–09h	0,5 h	45 s	≈42	4,1%	≈4,2%
09–10h	1 h	36 s	≈103	10,1%	≈10,3%
10–11h	1 h	38 s	≈95	9,7%	≈9,6%
11–12h	1 h	57 s	≈64	6,4%	≈6,4%
12–13h	1 h	53 s	≈68	6,9%	≈6,8%
13–14h	1 h	37 s	≈99	9,9%	≈9,9%
14–15h (pico)	1 h	**28 s**	**≈130**	**13,3%**	**≈13,0%**
15–16h	1 h	29 s	≈126	12,7%	≈12,6%
16–17h	1 h	34 s	≈107	10,6%	≈10,8%
17–18h	1 h	36 s	≈101	10,2%	≈10,1%
18–19h	1 h	59 s	≈61	6,1%	≈6,2%
Total do dia			**≈997**	**100%**	**100%**

Fonte: Elaborada pelo autor.

Para que, a cada execução, os clientes não cheguem exatamente com os mesmos espaçamentos de forma robótica, inclui-se um ruído de ±25%. Isso significa, por

exemplo, que na janela de pico, que possui espaçamento de 28s entre clientes, um cliente pode chegar com 21s a 35s de diferença para o último entrante. Dessa forma, dois dias simulados não compartilham os mesmos instantes de chegada.

A execução da simulação permite dois modos. No modo dia inteiro, simula-se um dia de loja normal. Os clientes começam a chegar a partir das 08:30 e param de entrar após as 19:00 (horário em que a loja começa a fechar). No modo pico, somente a janela das 14h até as 15h é simulada. Nesse intervalo, entram cerca de 130 clientes. Para comparação dos dois modelos (clássico e generativo) e execução dos experimentos com uso de LLM, optou-se pelo modo de pico, a fim de reduzir o custo computacional e financeiro, uma vez que os agentes realizam múltiplas chamadas ao LLM. Simular o dia inteiro multiplicaria o custo em aproximadamente sete vezes.

6.5 Agente Cliente

O agente cliente é o ator principal da simulação; ele é o consumidor, quem de fato faz os indicadores se moverem. Ele entra na loja, circula pelos setores avaliando os produtos, provando-os (quando possível) e decidindo se os compra. Ao final de seu tempo na loja, caso haja produtos em sua cesta, ele vai ao caixa; caso esteja vazia, ele vai direto à saída. No modo clássico, o seu fluxo de compra é baseado em regras fixas e premissas mecânicas; a sua persona existe e interfere no comportamento (por exemplo, definindo o orçamento e o tempo de permanência), ainda que o perfil sociodemográfico não guie as decisões de compra; enquanto no generativo a LLM é responsável pelas ações e decisões do agente com base na persona, memória e raciocínio.

6.5.1 Persona

Persona é o conjunto de características que dão origem à biografia dos agentes. Quando o cliente é criado, ele recebe estatisticamente cada uma das características citadas na Tabela 9, onde cada atributo possui uma fonte real. Sempre que possível, utilizou-se o dado mais atual e local disponível (Camaquã, depois Rio Grande do Sul, depois a região Sul e, por fim, o Brasil). Os atributos foram escolhidos de forma a descreverem da forma mais completa possível sem enviesar a compra de determinados itens.

Tabela 9 – Origem das características da persona do cliente.

Característica	**Distribuição / valor**	**Fonte (escopo)**
Gênero	69% mulher · 31% homem	IEMI, consumidor de vestuário (Brasil)
Renda per capita	10 décimos: R$233 a R$9.221/mês (abr/2026)	IBGE Síntese Ind. Sociais 2023, tab. 3 (PNAD 2022, Sul)
Idade	18+ pela pirâmide (18–39 ≈41% · 40–59 ≈35% · 60+ ≈24%)	IBGE Censo 2022, tab. 9514 (Camaquã)
Tem filhos	62% sim (quantidade 1–3 uniforme, premissa)	IBGE Censo 2022, tab. 2873÷9514 (RS)
Cor/raça	branca 78,4% · parda 14,7% · preta 6,5% · outras 0,4%	IBGE Censo 2022, tab. 9605 (RS)
Casa própria	73% sim	IBGE PNAD 2025, tab. 6821 (RS)
Estado civil	casado(a) 50% · solteiro(a) 35% · viúvo/divorciado 15%	IBGE Censo 2010, tab. 3193 (Camaquã)
Escolaridade	fundamental 44% · médio 36% · superior 20%	IBGE PNAD, tab. 4095 (RS)
Trabalha fora	63% sim	IBGE PNAD, tab. 4095 (RS)
Pratica esporte	24% sim	IBGE PNAD 2015, Supl. Esporte (RS)
Cliente recorrente	80% sim (frequentes)	Dado da loja (Camaquã)

Fonte: Elaborada pelo autor, a partir de tabelas do IBGE e do IEMI, além de dados da loja.

6.5.2 *Buyer* ou *Browser*

Em uma loja real, nem todo entrante compra. Foi decidido que, na simulação, o eixo *buyer*/*browser* seria ancorado em referências de conversão do varejo físico, o que foi confirmado pelo número de conversão real disponibilizado pela loja. Cerca de 2/3 das pessoas que entram não compram, apenas circulam pela loja, sem intenção de compra. Essa decisão de arquitetura foi tomada para facilitar o processo de avaliação dos compradores de fato. No modo clássico, os agentes observadores possuem trancas no fluxo de compra; no modo generativo, apenas os compradores de fato são atendidos pela LLM, enquanto os observadores permanecem clássicos com tranca.

6.5.3 Orçamento

O orçamento de cada cliente é o teto de gasto destinado àquela visita à loja; ele é calculado da seguinte forma:

$$O = R_{pc} \times V \times M \tag{1}$$

em que O é o orçamento, R_{pc} a renda per capita, V a fração gasta com vestuário e M os meses entre os gastos com vestuário.

A renda per capita vem da tabela de distribuição de renda domiciliar por décimos

da população (Tabela 12), disponível em (IBGE, 2023). Os valores que constam na tabela original eram escalados para a população brasileira do ano de 2022; logo, foi necessário reescalar para a Região Sul com o uso de uma tabela inclusa na mesma fonte, e atualizar para a data de abril de 2026 através da correção da inflação com o uso do IPCA, mais detalhes na Tabela 10. Na simulação, quando um cliente "nasce", é sorteado um percentil, que possui um valor de renda interpolado entre as médias dos décimos; esse valor é o R_{pc} utilizado no cálculo.

Tabela 10 – Fatores usados no cálculo do orçamento.

Uso	**Fator**	**Fonte**
Aplicados à renda (R_{pc}):		
Escala Brasil $\rightarrow$ Região Sul	$\times$1,215	(IBGE, 2023), Tab. 2
IPCA jun./2022 $\rightarrow$ abr./2026	$\times$1,177	(IBGE, 2026)
Usados para encaixar as classes da POF nos décimos:		
Despesa familiar $\rightarrow$ per capita (Sul)	$\div$2,64	IBGE, Censo 2022
IPCA 2018 $\rightarrow$ abr./2026	$\times$1,541	(IBGE, 2026)

Fonte: Elaborada pelo autor através dos dados do IBGE.

A fração da renda gasta com vestuário (V) é proveniente de um estudo de orçamento familiar (Lima; Araújo; Silva, 2024). Nele, foi medido o percentual de gasto de diversas categorias no gasto mensal das famílias. Na tabela, é perceptível a curva de Engel, fenômeno que descreve a diminuição da proporção de renda gasta com bens de necessidade à medida que a renda aumenta. Por ser uma divisão (gasto com vestuário sobre gasto total), V não se altera ao atualizarmos os valores absolutos; ainda assim, os valores de despesa total mensal presentes na tabela original foram trazidos para o recorte per capita e para reais de abril de 2026, dando origem aos valores da Tabela 11, apenas para que as classes de renda da POF pudessem ser encaixadas nos décimos do modelo. Então, para cada décimo, se encontrou o V, como mostra a Tabela 12.

Tabela 11 – Despesa com vestuário por classe de renda (POF 2018, recorte Sul; valores per capita em reais de abril de 2026).

Despesa total mensal (R$)	**Despesa com vestuário (R$)**	V **(%)**
960,14	36,77	3,8
1.413,83	52,91	3,7
2.182,09	86,87	4,0
3.552,63	141,50	4,0
5.319,99	192,15	3,6
9.138,05	258,04	2,8
14.411,04	317,25	2,2

Fonte: (Lima; Araújo; Silva, 2024), Tabela 10, recorte Sul. Valores originais (R$ de 2018) trazidos a per capita (÷2,64) e a reais de abril de 2026 (IPCA); V = despesa com vestuário ÷ despesa total, invariante à conversão.

Tabela 12 – Renda per capita e fração de vestuário por décimo (Região Sul, abril de 2026).

Décimo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Faixa (%)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
Percentil	5	15	25	35	45	55	65	75	85	95
R$/mês	233	542	782	1.004	1.274	1.603	1.919	2.497	3.604	9.221
V **(%)**	3,8	3,8	3,8	3,8	3,7	3,7	4,0	4,0	4,0	2,8

Fonte: renda de (IBGE, 2023) (Tab. 3); V de (Lima; Araújo; Silva, 2024) (Tab. 10, Sul), com cada décimo associado à sua classe de renda.

O intervalo médio entre gastos com vestuário possui como fonte uma pesquisa comercial da área (Opinion Box, 2024). No cálculo do orçamento, foi utilizada a mediana entre a quantidade de tempo respondido na pesquisa (2,5 meses). Duas observações merecem registro: a primeira é que os dados passam por conversões entre escalas (regional, temporal e per capita), todas com fonte declarada e detalhadas na Tabela 10. A segunda é que V é medida sobre o gasto mensal, enquanto no cálculo principal a multiplicamos pela renda mensal, o que acaba insinuando que toda renda é gasto. Essa é a realidade da maioria da população mais pobre, mas superestima o orçamento da cauda mais rica.

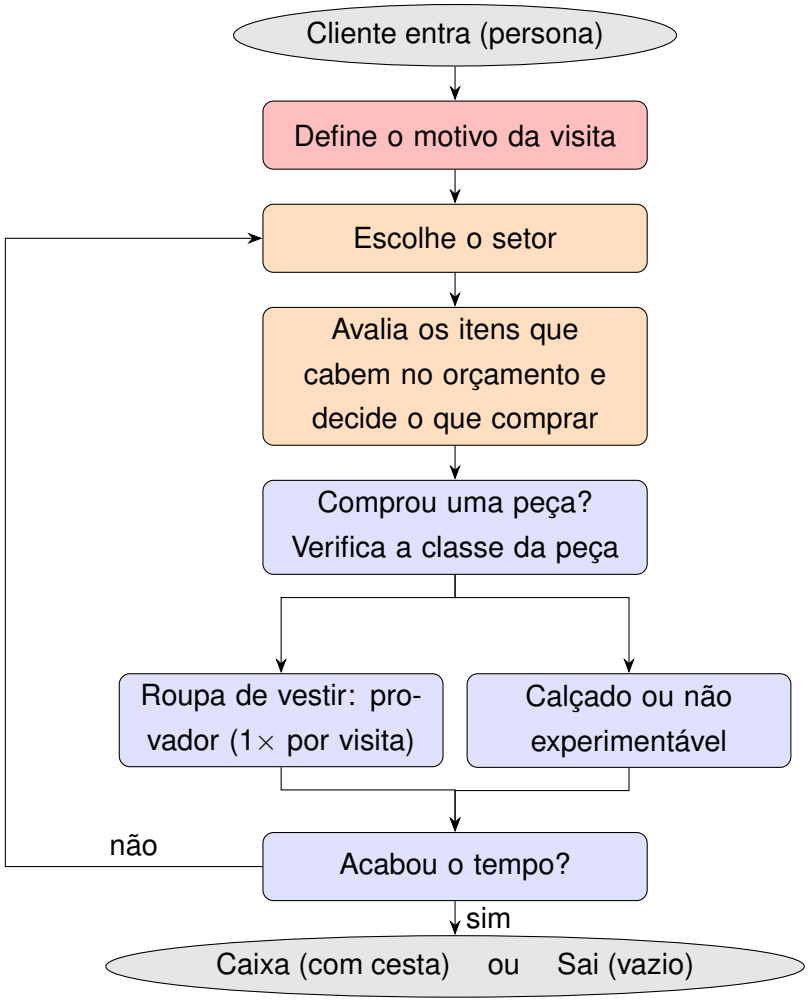
6.5.4 Funil de compra

O funil de compra é o conjunto de ações realizadas pelo cliente dentro da loja. O processo possui mecanismos iguais nas abordagens clássica e generativa, com exceção dos pontos de decisão, onde o cognitivo é acionado. Ao ser gerado, o cliente recebe orçamento, tempo de loja, persona e a informação de estar somente de visita (*browser*) ou de ser um comprador (*buyer*). A primeira decisão é: para qual setor ir. No modelo clássico, o setor é sorteado com probabilidade proporcional à sua área (Tabela 5), seguindo a premissa de que, quanto maior o setor, mais chance de uma

pessoa acabar visitando-o. Enquanto no modelo generativo, a LLM utiliza a sua persona para criar um motivo plausível para a visita e, a partir desse motivo e das demais características, o setor é escolhido. Ao chegar ao setor, a segunda decisão deve ser tomada: comprar ou não um produto daquele setor. No clássico, são sorteadas de 1 a 3 araras que o cliente avalia; cada uma dessas araras representa a avaliação de um item aleatório que cabe em seu orçamento, com chances iguais de comprar e de rejeitar. Caso rejeitado, o produto não aparece na próxima arara da mesma navegação. No generativo, uma avaliação ocorre da seguinte forma: a LLM recebe a lista de produtos com os preços reais e decide colocar um ou nenhum item na cesta; em seguida, pode avaliar novamente os itens do setor ou voltar a navegar. Em ambas as abordagens, cada avaliação acrescenta 90 segundos ao tempo de loja (no clássico, por arara; no generativo, por chamada). É importante citar que, enquanto o agente generativo espera o retorno de uma chamada da LLM, o seu tempo de loja fica congelado.

Com o item na cesta, a classe de avaliação do item será responsável pelas próximas ações do cliente. Caso o item seja uma roupa da classe “experimentável”, o cliente vai até o provador (um cliente vai apenas uma vez ao provador por visita); o provador adiciona 5 minutos e meio ao *dwell time*. O calçado e os itens não experimentáveis não acrescentam tempo além da própria avaliação do item, pois, apesar de se saber que provar calçado leva mais tempo, optou-se por não incluir mais uma premissa sem fonte para a magnitude. Em seguida, o cliente é direcionado ao último bloco antes de recomeçar o fluxo, onde é verificado se o seu *dwell time* já chegou ao limite. Caso tenha chegado e ele tenha itens na cesta, vai ao caixa finalizar a compra; caso tenha chegado e esteja de cesta vazia, direciona-se à saída; caso o cliente ainda tenha tempo de loja sorteado no início, reinicia o fluxo de compra com o processo de navegação. É importante citar que, durante o funil de compra, o cliente pode ser abordado por um agente vendedor. No modo clássico, a conversa apenas adiciona entre 15 e 25 s ao seu *dwell time*; no modo generativo o tempo também é contabilizado, entretanto, ocorre de fato a conversa entre os agentes, que é armazenada na memória do agente cliente e influencia decisões futuras. O fluxo se encontra na Figura 6, cujos blocos se dividem da seguinte forma: em cinza, o início e o fim do fluxo; em vermelho, a definição da história (o motivo da visita), que ocorre só no generativo; em laranja, os dois pontos cujo modo de decisão muda entre as abordagens (a escolha do setor e a avaliação da peça: sorteio no clássico, LLM no generativo); e, em azul, as etapas físicas idênticas nos dois modos (o roteamento por classe, o provador e o controle de tempo).

Figura 6 – Funil de compra do agente cliente.



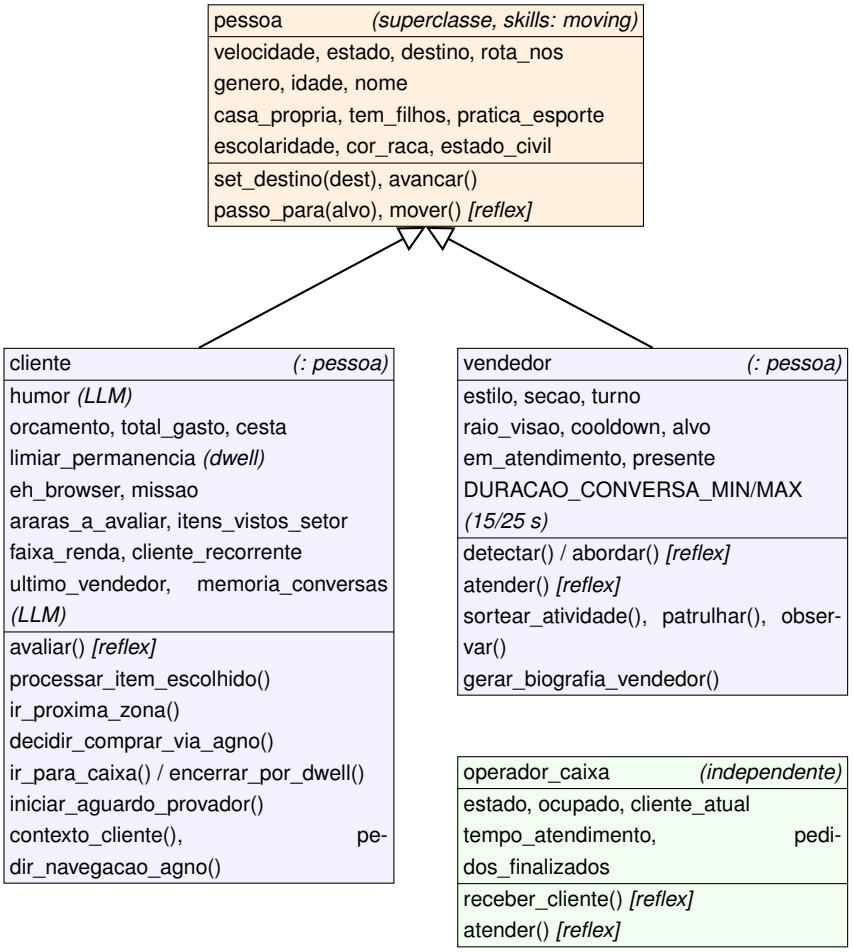
Fonte: Elaborada pelo autor.

6.6 Agente Vendedor

Apesar de não influenciar diretamente os indicadores, o agente vendedor é capaz de interferir no resultado de vendas de uma loja através do seu contato com o cliente. A pesquisa (Mindtree, 2016) indica que a interação cliente-vendedor pode aumentar a chance e a quantidade da venda de produtos. No clássico, essa interferência acaba sendo apenas temporal, devido ao tempo de duração das abordagens; não foi implementado o efeito de persuasão de forma mecânica, pois é algo que deve ser contextual. Já no generativo, a abordagem gera uma conversa em que ambos sabem o seu papel: o vendedor possui acesso aos itens que o cliente possui em sua cesta (para poder comentar sobre eles), mas não há de fato um script do que dizer ou de como agir. Como resultado, o agente cliente leva a opinião da conversa para a sua memória. Forma-se uma conexão entre cliente e vendedor após a sua primeira abordagem, evitando que outros vendedores interajam com o cliente e mimetizando o processo que acontece na loja real (“Já foi atendido(a)?” / “Sim, obrigado(a).”). Caso

o vendedor reaborde o cliente, ambos possuem acesso ao que aconteceu na última conversa.

Figura 7 – Hierarquia de classes dos agentes.



Fonte: Elaborada pelo autor.

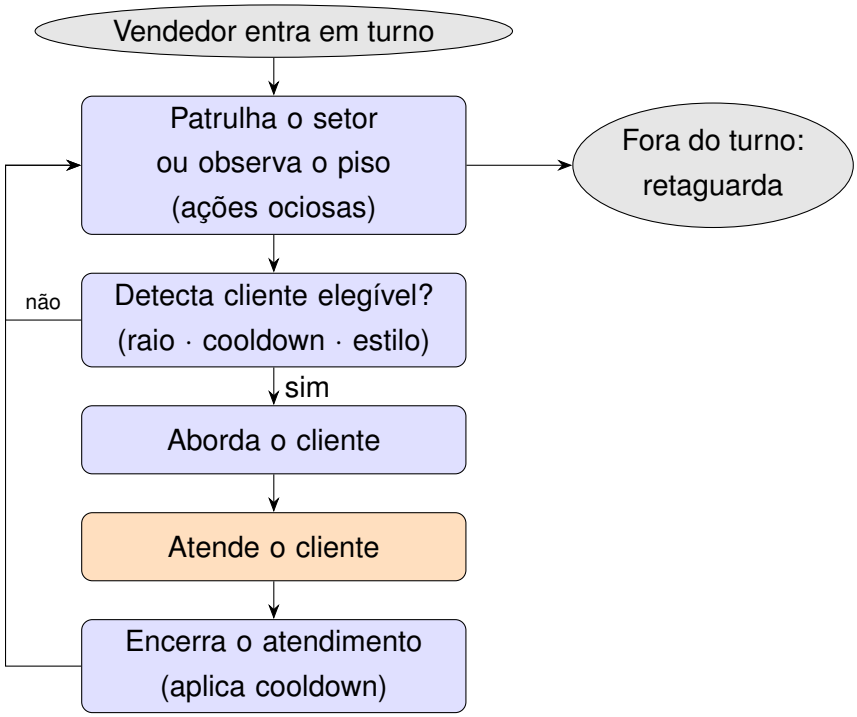
Como apresentado na Figura 7, o agente vendedor possui atributos que, assim como no agente cliente, são herdados de uma classe pai chamada pessoa; mas as duas filhas acabam diferentes, pois cada uma tem atributos e métodos específicos. No caso do vendedor, por exemplo: estilo, a mecânica que define a situação em que ele aborda o cliente (Farese; Kimbrell; Woloszyk, 2012); raio de visão, premissa que indica a distância em que o cliente deve estar do vendedor para ser abordado; cooldown, premissa de tempo mínimo entre duas abordagens do mesmo cliente; alvo, o cliente que está sendo abordado; e a duração da conversa (de 15 a 25 s), premissa responsável pelo tempo que a abordagem consome do cliente. Entre os métodos específicos estão o de detectar e abordar um cliente próximo, e o de retornar à sua posição ao encerrar o atendimento.

O operador de caixa, também presente na Figura 7, foi modelado como uma classe independente, que não herda de pessoa. Diferentemente do cliente e do vendedor, ele não se desloca pela loja (a superclasse pessoa carrega justamente a habilidade

de movimento e os atributos sociodemográficos) e atua como um ponto fixo de atendimento, à semelhança do provador. Herdar de pessoa apenas acrescentaria atributos que nunca seriam utilizados; por isso, o operador mantém somente o estado necessário para processar a fila do caixa (se está ocupado, o cliente em atendimento e o tempo de processamento), cujo funcionamento é detalhado na seção de filas e atendimento.

A equipe de vendedores da loja foi modelada a partir da loja real: é constituída por 18 vendedores (3 masculinos e 15 femininos), divididos em 2 turnos. O turno A possui 12 vendedores, que trabalham das 08h às 18h (com almoço das 11h às 13h); o turno B possui 6 vendedores, em meio-período, das 09h às 13h, sobrepondo-se ao turno A das 09h às 11h (o pico de equipe da manhã). Além da abordagem, o vendedor possui um funil de ações ociosas que se resume a patrulhar o setor e observar o piso, permanecendo fixo na sua seção. O funil de ações do vendedor se encontra de forma completa na Figura 8, cujos blocos se dividem da seguinte forma: em cinza, o início e o fim do turno do vendedor; em azul, as etapas mecânicas, idênticas nas duas abordagens; e, em laranja, o único ponto que muda entre elas: a *conversa*.

Figura 8 – Funil de ações do agente vendedor.



Fonte: Elaborada pelo autor.

6.7 Filas e atendimento

As filas são sintomas indicativos de gargalos no fluxo e podem ser formadas através do uso do provador e do caixa. O número de provadores e caixas disponíveis é baseado na loja real: os provadores são 9 (6 femininos e 3 masculinos) e a quantidade

de caixas é 10, porém, a quantidade de caixas ativos é parâmetro com valor padrão 5, pois é um valor que pode mudar a depender da demanda. A quantidade de tempo de prova no provador tem base em benchmark do setor (V-Count, 2023), enquanto a duração do atendimento no caixa é uma premissa de valor 2 minutos, que aumenta 30 s por item na cesta.

Ambas as mecânicas de caixa e provador possuem seus pontos de espera, como mostra a Figura 5, e suas filas seguem o princípio FIFO. O provador conta com 6 pontos de fila para cada gênero (feminino e masculino) e o caixa, com 5 pontos sem gênero. Vale citar que um ponto pode abrigar mais de um cliente quando os demais estão ocupados. Um comportamento presente nas lojas, mas que não foi modelado, é o *walk-off* de filas, que, apesar da falta de magnitude para implementação na abordagem clássica, poderia ter sido incluído e medido na parte generativa.

6.8 LLM e cognição

Esta seção detalha como a camada cognitiva da abordagem generativa foi implementada no projeto, permitindo que as decisões sejam resultado do raciocínio da LLM a partir do contexto, e evitando a necessidade de premissas com números estimados. Encontra-se dividida em três subseções. A primeira detalha como acontece a conexão entre o motor cognitivo (Agno + LLM) e o motor espacial e de regras (GAMA); a segunda trata dos momentos em que a cognição entra de fato em ação; e a terceira aprofunda a questão da persona e como funciona sua memória.

6.8.1 Arquitetura da ponte GAMA–Agno

A ponte é constituída de duas estruturas intermediárias: a primeira é um processo local que roda paralelamente à simulação. Esse processo é escrito em Python e chamado de sidecar. Ele fica verificando periodicamente (*poll*) uma pasta pré-definida compartilhada com o GAMA. Quando um agente ativa um gatilho de decisão, o GAMA escreve a requisição na pasta em formato JSON e passa a fazer poll à espera da resposta. A requisição contém o contexto atual daquele agente, e é nomeada no formato `ponte_req_idcliente_nonce`, diferenciando-a através do id único de cada cliente somado ao nonce (número do ciclo em que a requisição foi realizada). O sidecar visualiza a requisição e envia o pedido à segunda estrutura existente na ponte, a API Agentic MAS (Alves; Born; Aguiar, 2026), que é responsável por orquestrar as chamadas cognitivas. A API encaminha o contexto ao Agno, que monta o agente correspondente com sua persona e aciona a LLM; o modelo retorna a resposta em JSON estruturado para a API, que retorna para o sidecar.

De posse da resposta, o sidecar cria um arquivo do tipo `ponte_resp_idcliente_nonce` na pasta que o GAMA está observando. Esse

arquivo, quando pareado com sua requisição, torna a existência da mesma desnecessária, que é então removida da pasta, mantendo apenas a resposta para fins de análise. O GAMA lê a resposta e aplica a decisão da LLM no fluxo do agente, fechando o ciclo. Através desse método, todo agente consegue iniciar seu gatilho de forma assíncrona, sem a necessidade de congelar a simulação, acelerando o processo, mas ao custo de perder precisão no relógio da simulação.

A API Agentic MAS é de propósito geral, possibilitando atender qualquer tipo de simulação multiagente, mas justamente por isso não dá profundidade a nenhum contexto específico, que no caso do projeto é a simulação de uma loja de vestuário varejista. Para usá-la no contexto do projeto, foi necessário realizar algumas adaptações para que funcionasse como planejado. Ela é uma API REST que, a cada requisição, recebe o contexto de um agente e devolve a decisão correspondente estruturada em JSON. A adaptação concentrou-se na rota *decisions*, responsável por acionar a LLM: em sua forma original, ela produzia uma única decisão genérica, a escolha de uma ação abstrata como mover ou aguardar; no projeto, foi especializada nos quatro gatilhos de decisão que a simulação exige, cada um com seu próprio contexto e formato de resposta esperado, a saber, história, navegação, conversa e item. Esses quatro gatilhos serão detalhados na próxima subseção.

Além da rota responsável pelas decisões, a arquitetura original da API propõe outras três rotas (*simulations*, *rounds* e *agents*), responsáveis pela persistência dos dados no banco; o que cada uma armazena já foi apresentado na Seção 3.3. No presente projeto, essas rotas foram mantidas, e a cópia que elas persistem no banco funciona como um *fallback*: a análise se apoia no módulo de coleta de dados descrito na Seção 7.2, mas esses dados poderiam ser recuperados do banco caso houvesse alguma falha nesse módulo. A rota *rounds* é chamada uma única vez no início da execução, pois toda a simulação corresponde a um único período de simulação (dia inteiro ou horário de pico), e não a uma sucessão de períodos, como diferentes dias, para a qual a segmentação em múltiplas rodadas foi projetada. Outro ponto de diferenciação diz respeito à memória agêntica durante a visita: foram testados dois métodos, a persistência gerenciada pelo próprio banco e o envio da memória como parâmetro no contexto da chamada. Ambos funcionaram adequadamente para a necessidade do projeto, mas, como não foram simuladas as múltiplas visitas à loja de um mesmo cliente (caso em que a rota *rounds* seria utilizada), optou-se pelo método mais simples, o de envio no contexto.

6.8.2 As decisões cognitivas e seu contexto

Durante o fluxo de compra do cliente, há diversos momentos em que existe a necessidade de tomar decisões que, no modelo clássico, são resultado de regras e sorteios, mas, na abordagem generativa, são delegadas à LLM. Estes são os quatro

momentos do fluxo do cliente em que há um gatilho de chamada à LLM:

- **História**: Quando um cliente é gerado, ele deve gerar um motivo para a sua ida à loja com base no contexto que recebe
- **Navegação**: A escolha de para qual setor se dirigir em seguida, dentre os departamentos disponíveis.
- **Escolha de item**: durante a avaliação em um setor, a decisão de colocar uma peça na cesta, ou nenhuma, selecionando entre os itens candidatos que cabem no orçamento.
- **Conversa**: o diálogo com o vendedor, que também é um agente generativo, conduzido em múltiplos turnos até que o cliente o encerre. O comportamento dessa interação é detalhado na seção do Agente Vendedor.

Durante as quatro requisições, a LLM recebe o contexto necessário para a decisão, montado pelo GAMA e enviado na requisição. Todas partem de um mesmo contexto-base (Tabela 13), que reúne a persona do cliente, a descrição da loja (uma loja de vestuário popular em Camaquã) e o mês dos dados (abril), além da memória da visita: humor, cesta de itens, departamentos atual e já visitados, orçamento restante e os pensamentos das conversas que teve. Sobre essa base, cada requisição acrescenta o que a decisão exige. Na história, gerada na chegada, a memória ainda está vazia. Na navegação, somam-se os setores disponíveis; na escolha de item, a lista de produtos do departamento avaliado que cabem no orçamento restante do cliente, com seu nome e preço originais. A conversa é a situação que mais se distingue das outras, ela envolve um segundo agente generativo, que é gerado através de uma persona própria somada ao tom de abordagem (O padrão é neutro); o cliente não tem acesso a essas informações, percebendo apenas o que o agente lhe é dito, já o vendedor possui acesso a cesta do cliente e os produtos do catálogo de produtos do setor que ambos se encontra, assim conseguindo gerar conversas sobre o que vê e tentativas de persuasão para compra.

Tabela 13 – Contexto-base enviado à LLM em toda decisão.

Campo	Conteúdo
Persona (estável na visita)	
`persona`	Nome, gênero, idade, estado civil, filhos, escolaridade, cor/raça, moradia, trabalho, esporte, faixa de renda e se é cliente recorrente
Loja e data	
`loja`	Lojas ***** (Camaquã/RS), loja popular de moda e casa
`data`	abril de 2026 (mês dos dados)
Estado e memória da visita (variável)	
`motivo_visita`	A história (motivo da visita) gerada na chegada
`setor_atual`	O departamento em que o cliente se encontra
`humor`	Estado emocional atual
`ja_visitou`	Departamentos já visitados
`cesta`	Itens já colocados na cesta (por nome) e a contagem
`orcamento_restante`	Valor ainda disponível na carteira
`conversas_com_vendedor`	Pensamentos guardados das conversas anteriores

Fonte: Elaborada pelo autor.

É importante ressaltar que os prompts de cada requisição foram desenvolvidos de maneira a não enviesar o modelo na tomada de decisões, sem ponderamento de características, de modo ao resultado emergir da combinação entre persona, situação, memória e raciocínio da própria LLM. No entanto, para evitar alucinações foram adicionadas algumas frases de maneira a filtrar erros como mostrados na seção 7.3.5. O template de prompt de cada uma das decisões se encontra no apêndice A.

Independentemente do tipo, a resposta retorna sempre em formato estruturado, em JSON, que o GAMA aplica de forma determinística: o setor escolhido vira o próximo destino, e o nome do item retorna ao produto real correspondente na simulação. Essa separação preserva a disciplina do modelo: na conversa, por exemplo, quando o cliente decide levar uma peça, não cabe à LLM calcular se ela cabe no orçamento nem fixar o preço; o GAMA confere o orçamento de forma mecânica e resolve o nome informado no produto real. A decisão cognitiva, querer a peça, fica assim separada da verificação determinística, poder comprá-la. Quando a resposta é inválida ou ambígua, um tratamento de robustez a resolve sem enviesar o resultado, por exemplo escolhendo de forma uniforme entre os candidatos válidos.

6.8.3 Persona determinística e memória

A persona, cujos atributos e fontes foram apresentados na Tabela 9, na seção do cliente, é responsável por introduzir a distribuição de orçamento da população e o eixo *browser*/*buyer*, que aciona a trava do fluxo de compra no caso de um browser; ambas

as características são utilizadas nas abordagens generativa e clássica. Os demais atributos, sorteados de forma probabilística, compõem a biografia dos agentes, inserida no campo `description` do Agno, que acompanha e influencia todas as chamadas à LLM.

O sorteio das características gera um conjunto de personas determinístico, sendo possível reproduzir a mesma população ao repetir a seed. Essa decisão foi tomada a fim de garantir a justiça na comparação entre as abordagens generativa e clássica, uma vez que ambas as execuções teriam o mesmo orçamento total para gastar e a mesma quantidade de compradores. Dessa forma, qualquer diferença observada nos resultados decorre do mecanismo de decisão, regras contra LLM, e não de populações distintas.

Como se mostrou na seção anterior, optou-se por manter a memória do cliente inteiramente em tempo de execução, em atributos do próprio agente no GAMA; a cada nova decisão, todo esse conteúdo é inserido no contexto da requisição à LLM, de modo que ela decide com plena consciência de por onde o cliente já passou, do que já carrega, do quanto ainda pode gastar e da experiência acumulada até ali. Convém distinguir os dois conceitos: a persona é um input estável, responsável pela heterogeneidade da população, enquanto a memória é acumulada dinamicamente ao longo da visita. Esse conteúdo, tanto os pensamentos quanto os diálogos, é o que fica registrado no transcript descrito na Seção 7.2.

A continuidade da conversa é tratada à parte. Ela não reside no GAMA: é orquestrada pelo *runner*, o componente da API que constrói o agente no Agno, aciona a LLM e conduz a conversa. É o runner que mantém em memória, apenas enquanto a simulação roda, o histórico da última conversa entre cada par cliente-vendedor. Assim, quando o mesmo vendedor reaborda um cliente que já havia atendido, o runner recupera esse histórico e o fornece ao prompt, preservando a continuidade do atendimento em vez de recomeçar do zero. Esse tipo de memória também existe em tempo de execução, mas é gerenciado pela API e não pelo GAMA.

7 ANÁLISE DOS RESULTADOS

Este capítulo apresenta os resultados: as medidas de custo, latência e robustez que viabilizam a abordagem generativa, a coleta de dados, o experimento principal que trata da análise comparativa entre as abordagens clássica e generativa na reprodução do comportamento da loja real, e os três experimentos contrafactuais que aplicam estratégias de gestão e verificam os resultados emergentes (layout, promoção e atendimento).

7.1 Custo, latência e robustez

A abordagem generativa tem um custo financeiro e temporal alto, cada decisão exige no mínimo uma requisição ao modelo de LLM, ao longo de uma conversa entre cliente e vendedor, os agentes podem gerar até 14 chamadas de LLM. Para viabilizar os experimentos utiliza-se duas estratégias: A primeira trata-se da diminuição do período simulado, ao invés de o fluxo de um dia inteiro de loja, restringe-se ao fluxo de clientes do horário de pico (14h – 15h). A segunda é que apenas a parte compradora dos clientes que entram na loja utiliza a LLM (1/3), enquanto a parte visitante faz uso da cognição clássica. Essas restrições são decisivas para a viabilidade, pois estender as chamadas à LLM a toda a jornada diária multiplicaria o custo por aproximadamente sete vezes, sem alterar o fenômeno que se deseja observar, que é a decisão de compra.

O modelo de linguagem escolhido foi o *gpt-4o-mini*, da OpenAI, que consegue conciliar baixo custo, velocidade, flexibilidade diante do fluxo massivo de chamadas simultâneas, capacidade de inferência necessária e limites de requisição elevados. A Tabela 14 apresenta os custos ao simular o fluxo de clientes no horário de pico. Para dimensionar o custo temporal introduzido pela camada cognitiva, vale o contraste com a abordagem clássica: a mesma coorte de pico, sem nenhuma chamada à LLM, é processada em cerca de 26 segundos, ante os aproximadamente 22 minutos da versão generativa. A avaliação comparativa de modelos posicionados em diferentes pontos do *trade-off* entre custo/latência de inferência e capacidade de raciocínio fica

como proposta para trabalhos futuros.

Tabela 14 – Tempo de execução e custo por run no modo de pico.

Seed	Ciclos	Tempo (min)	Custo (US$)
42	—	—	0,22
1	6.848	23,05	0,22
2	6.562	22,09	0,20
3	7.035	23,68	0,23
5	6.790	22,86	0,23
7	6.585	22,17	0,22
11	6.744	22,70	0,20
99	6.436	21,67	0,25
Média	**6.714**	**22,60**	**0,22**

Fonte: Elaborada pelo autor.

Apesar da grande velocidade de resposta do modelo escolhido, as chamadas à LLM introduzem uma latência inevitável; logo, o *dwell time* seria afetado de forma injusta pelo tempo em que o cliente fica esperando o resultado da decisão. Para evitar essa contaminação pelo tempo de espera, introduziu-se o mecanismo *dwell-pause* (congelamento do relógio interno do cliente). Dessa forma, o tempo de permanência reflete apenas o período em que o cliente está efetivamente ativo na loja (navegando, avaliando ou conversando), e não a latência da infraestrutura, o que mantém a métrica comparável entre as abordagens clássica e generativa.

Por fim, a abordagem generativa depende de um serviço externo; em caso de manutenção ou instabilidade, um único erro poderia corromper uma análise inteira. A fim de prevenir esse tipo de ocorrência, criou-se um mecanismo de *fallback*: diante de um erro transitório na requisição, ela é reenviada automaticamente, em até cinco tentativas com espera crescente entre elas; se ainda assim o erro persistir, o fluxo generativo recorre ao fluxo clássico. Nas execuções que compõem os experimentos deste trabalho, no entanto, não houve nenhuma ocorrência de *fallback* para o fluxo clássico. Ainda assim, caso ocorresse em poucas decisões, o efeito diluir-se-ia e não comprometeria a análise como um todo.

7.2 Coleta de dados

O resultado de cada execução é coletado e documentado para análise de duas formas. A primeira é a geração do transcript, um arquivo produzido a cada execução pelo sidecar enquanto lê as requisições e respostas na pasta da ponte compartilhada com o GAMA. A partir dessas chamadas e respostas, ele gera um arquivo do tipo JSONL, no qual cada linha é um JSON cujos dados dependem do tipo de decisão tomada pela LLM, como mostra a Tabela 15, tornando possível visualizar o histórico de decisões

de cada cliente na abordagem generativa. A segunda é a família de arquivos CSV (*Comma-Separated Values*) produzidos pelo GAMA em tempo de execução, detalhados na Tabela 16. Os quatro primeiros acumulam os resultados de todas as execuções em um mesmo arquivo, diferenciados pela seed e pela `run_tag` (identificação da execução) gravadas como primeiras colunas; o heatmap, por sua vez, é gravado como um arquivo por execução, com a seed e a `run_tag` no próprio nome.

Tabela 15 – Registro cognitivo armazenado no transcript

Tipo de decisão	**Conteúdo registrado**
`historia`	O motivo da visita gerado na chegada
`navegacao`	O setor escolhido como próximo destino
`escolha_item`	O item avaliado (levado ou recusado) e o humor resultante
`conversa`	O pensamento e o humor resultantes e o diálogo completo com o vendedor

Fonte: Elaborada pelo autor.

Tabela 16 – Arquivos CSV gerados pelo GAMA com o desfecho da simulação

Arquivo	**Granularidade**	**Conteúdo**
`kpi_runs`	Por execução	Vendas, ticket, itens por cupom, conversão, composição do mix por departamento e coorte de pico
`clientes_run`	Por cliente	Grupo (clássico/generativo), tipo (comprador/observador), perfil (gênero, idade, renda), missão, se comprou, itens, gasto, orçamento, tempo e motivo de saída
`dwell_por_agente`	Por cliente	Tempo total e permanência em cada zona da loja
`itens_vendidos`	Por produto	Nome, departamento, quantidade e valor efetivamente vendidos
`heatmap`	Por célula	Posição na grade e ocupação, separada entre compradores e observadores

Fonte: Elaborada pelo autor.

7.3 Experimento principal

Essa seção apresenta os resultados do experimento principal, respondendo à seguinte pergunta: qual das duas abordagens representa melhor a loja real? A comparação utilizou a média de oito execuções diferentes do horário de pico; ambas as abordagens (clássica e generativa) utilizaram as mesmas *seeds*, mantendo a população e diferenciando-se apenas na parte cognitiva.

7.3.1 Conversão, Ticket, Cesta e Faturamento

A conversão, apesar de ancorada nos valores reais, diferencia-se entre as abordagens: na abordagem generativa, um agente cliente que deveria ser comprador pode acabar decidindo sair sem comprar nada, um caso relativamente raro, mas que, dos 352 clientes compradores nessas 8 *seeds*, ocorreu 38 vezes (10,8%). No clássico, isso não acontece, pois não há o poder de decisão; o que pode ocorrer é o tempo de loja não ser suficiente para o cliente sortear e comprar um item que caiba no seu orçamento, e ainda assim apenas 4 (1,2%) clientes saem de cesta vazia. Quanto ao ticket médio, o generativo ficou abaixo da mediana (R$171) e o clássico ligeiramente acima (R$185), mas ambos ficaram abaixo da média (R$229), evidenciando que a cauda dos mais ricos (*tickets* altos que puxariam a média para cima) não foi bem representada pelo cálculo de orçamento atual.

Tabela 17 – Comparação dos KPIs agregados entre abordagens e a realidade.

KPI	Clássico	Generativo	Real
Conversão (%)	32,3	29,8	31,8
Ticket médio (R$)	185,3	171,0	229 (média) / 178 (mediana)
Itens por cupom	2,39	2,67	2,59
Faturamento (R$)	7.900	6.705	≈9.500

Fonte: Elaborada pelo autor.

A abordagem generativa chegou mais próxima à quantidade de itens esperados na cesta de forma emergente (2,67 contra 2,59), enquanto a clássica, nos valores de premissa calibrados, fica ligeiramente abaixo (2,39). Cabe notar que essa diferença não é uma limitação intrínseca do clássico: a análise de inversão (Tabela 18) mostra que aumentar o tempo de permanência não basta, pois a cesta satura em torno de 2,53, limitada pelo orçamento; reduzir o tempo de avaliação das araras aproxima a cesta de 2,59, mas ela também satura, em torno de 2,55 no piso plausível da faixa, de modo que alcançar o valor cheio exigiria um tempo de arara implausivelmente curto ou elevar a probabilidade de compra. Não se adotou nenhum dos dois: o tempo de avaliação é compartilhado pelas duas abordagens, sendo injusto alterá-lo apenas no clássico, e a probabilidade de compra só atingiria o alvo acima da banda de máxima entropia que a premissa sustenta, configurando *overfitting*. Quanto ao faturamento, ambas as abordagens ficaram abaixo da estimativa para o horário de pico, indicando novamente a falta da cauda rica. Através desses números, pode-se verificar que ambas as abordagens se aproximaram de forma consistente dos números agregados da loja, mas isso não quer dizer que a loja esteja sendo bem representada por completo, pois é possível chegar aos números certos pelas maneiras erradas.

Tabela 18 – Análise de inversão: valor que cada premissa precisaria assumir, isoladamente, para o clássico atingir a cesta real (2,59 itens por cupom).

Premissa	**Escopo**	**Calibrado**	**Para 2,59**	**Situação**
Tempo de permanência	compartilhado	36 min	satura em $\approx$2,53	não atinge 2,59: a cesta satura, limitada pelo orçamento
Tempo de avaliação/arara	compartilhado	90 s	satura em $\approx$2,55	no piso plausível (40 s da faixa 40–120 s) a cesta chega a $\approx$2,55; 2,59 exigiria arara $<$15 s
Probabilidade de compra	só clássico	0,50	$\approx$0,78	fora da banda de máxima entropia (0,40–0,65)

Fonte: Elaborada pelo autor.

7.3.2 Mix por departamento

Esse indicador divide o faturamento da loja entre os departamentos, permitindo visualizar qual abordagem representou melhor a distribuição da demanda dos clientes entre as seções. É a métrica em que as duas abordagens mais se distinguem e, ao contrário do esperado, o clássico vence de forma clara.

A Tabela 19 apresenta o mix de departamentos pelo valor vendido, indicando qual é a fatia de cada departamento no faturamento total da loja. O clássico possui um erro de representação de apenas 13,7 pontos, e o mix que produz é bastante fiel, com desvios relevantes apenas no infantil (subatendido, 17,8 contra 23,3) e no calçados (superatendido, 21,1 contra 16,7). O generativo, por outro lado, infla o feminino (que de fato é o setor mais vendido), mas abandona de forma extrema o infantil, subrepresenta o masculino e mostra um interesse além do necessário no departamento de cama, mesa e banho.

Tabela 19 – Mix por departamento (% do valor vendido).

Departamento	**Real**	**Generativo**	**Clássico**	**Erro Gen.**	**Erro Cláss.**
Feminino	31,6	36,9	30,4	5,3	1,2
Masculino	20,4	14,7	21,5	5,7	1,1
Infantil	23,3	12,6	17,8	10,7	5,5
Calçados	16,7	19,8	21,1	3,1	4,4
Cameba	5,8	13,2	7,2	7,4	1,4
Beleza	2,1	3,0	2,0	0,9	0,1
Erro absoluto				**33,1**	**13,7**

Fonte: Elaborada pelo autor. Real conforme `VendasPorSetor` (abril). Erro Gen. e Erro Cláss.: diferença absoluta de cada abordagem em relação ao Real, em pontos percentuais.

O resultado contém duas ressalvas. A primeira é metodológica: optou-se por ponderar a navegação do clássico com base na área de cada departamento, em vez de

sorteá-la de forma totalmente aleatória, o que deixaria o modelo sem qualquer ancoragem, servindo como um espantalho. Essa escolha tem respaldo empírico, pois estudos clássicos já demonstraram a correlação positiva entre o espaço de prateleira e as vendas (Curhan, 1972). A navegação do generativo, por outro lado, parte da decisão da LLM, que é bastante sensível à engenharia de *prompt*, o que sugere haver espaço para enriquecer a biografia das personas.

A segunda é uma limitação da camada cognitiva, associada à ausência de dinâmica familiar no modelo: cada agente é simulado como um indivíduo isolado, que decide e compra para si, e não como parte de um núcleo familiar. Apesar de 62% da população gerada ter filhos, a LLM ainda sub-representa a compra no departamento infantil, e essa limitação se manifesta de duas formas. Primeiro, o agente priorizou a compra de produtos para si, em vez de adquiri-los para outra pessoa (filhos, netos, sobrinhos, etc.), casos que existem, mas não são recorrentes. Segundo, não há simulação da compra em grupo, situação em que esse departamento provavelmente ganharia vida, como uma mãe indo com os filhos ou avós indo com os netos. Boa parte da demanda real do infantil nasce justamente dessa dinâmica familiar, o provisionamento do lar e a criança que acompanha e influencia a compra, que o modelo, centrado no agente individual, não representa.

7.3.3 Ranking de itens

A análise de ranking de itens compara os itens mais comprados na loja real com os comprados pelos clientes na simulação com as abordagens clássica e generativa. É importante relembrar que a abordagem clássica possui navegação ponderada, mas a decisão de compra é totalmente aleatória dentre os itens que cabem em seu orçamento. A Tabela 20 mostra a quantidade de itens comprados em diferentes intervalos de ranqueamento. Ambas abordagens tiveram resultados muito parecidos, mas de natureza completamente distinta, enquanto no clássico os itens variados e baratos reinam (*kits* de meia, camisetas), o generativo parece perceber o contexto de temperatura, e compras de casa, focando fortemente em blusas (`BLUSA MANGA LONGA`, `BLUSA CONTEMPORANEA`). A Tabela 21 detalha os itens mais comprados por unidade em cada abordagem, evidenciando esse contraste de perfil. O catálogo real possui 209 famílias, das quais o generativo deixou cerca de 97 sem comprar, enquanto o clássico cobriu a grande maioria, deixando apenas cerca de 33 de fora e de forma mais equilibrada.

Tabela 20 – Itens do top-N por quantidade de cada abordagem que constam no top-N real (também por quantidade de unidades vendidas).

Abordagem	Top 10	Top 20	Top 30
Generativo	1 (10%)	6 (30%)	9 (30%)
Clássico	3 (30%)	4 (20%)	9 (30%)

Fonte: Elaborada pelo autor. Valores reais (abril).

Tabela 21 – Dez itens mais comprados por unidade em cada abordagem (soma das 8 *seeds*).

Generativo		Clássico	
Item	**Un.**	**Item**	**Un.**
Blusa manga longa adulto	112	Meia bebê	49
Toalha rosto	55	Tanga e biquíni	46
Joias e bijuterias	39	Calcinha alta	34
Toalha banho	34	Kit meia feminina	26
Blusa contemporânea	33	Kit meia masculino	22
Camiseta manga longa adulto inverno	29	Kit meia inf./juv. menina	18
Camiseta básica	27	Kit meia inf./juv. menino	17
Lençol avulso casal	27	Toalha rosto	13
Chinelo de inverno feminino	24	Fronha	13
Blusa MGA 3/4 adulto	21	Kit cueca adulto	12

Fonte: Elaborada pelo autor.

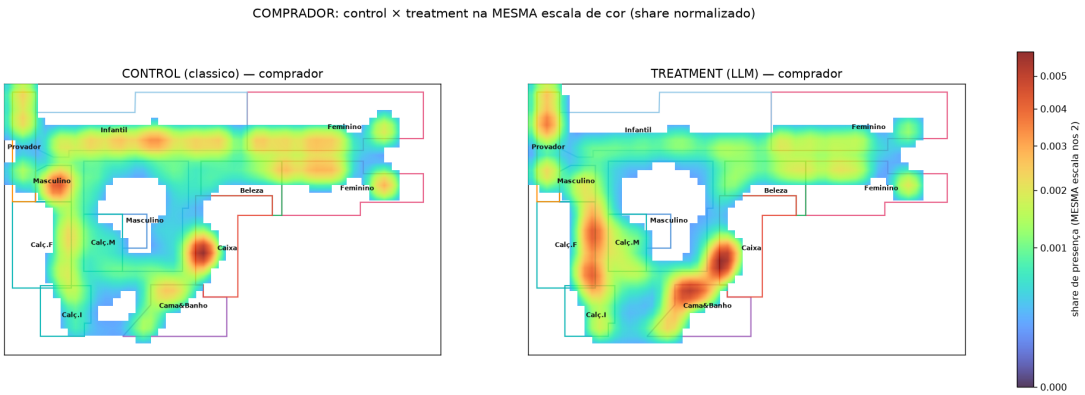
7.3.4 Dinâmica espacial

Além do quanto e do que se compra na simulação, a forma como o cliente circula pode gerar diagnósticos importantes. Pela Figura 9, percebe-se que ambas as abordagens não ”preenchem” todo o chão de loja; isso se dá devido à decisão metodológica de posicionar os pontos de avaliação nas bordas dos departamentos, evitando o cruzamento por fora do layout. Do lado esquerdo, verifica-se que os clientes da abordagem clássica se distribuem de forma mais homogênea entre os setores, com amarelados fortes em roupas femininas e infantis e um ponto avermelhado na roupa masculina que, apesar de corresponder ao terceiro departamento de maior área, pode gerar confusão: sua cor mais escura não representa maior circulação que os outros dois, mas sim a proximidade dos pontos de avaliação, devido à pouca área da zona masculina esquerda.

Ao analisarmos a camada cognitiva (lado direito), percebem-se alguns comportamentos. O maior fluxo nos provadores (10,4% da ocupação dos compradores, contra 7,0% no clássico) indica que o generativo concentra as compras em peças que passam pela prova, sobretudo blusas, enquanto o clássico, ao comprar de forma uniforme, distribui-se por itens variados, incluindo os *kits* (meia, cueca, calcinha) que não levam ao provador. Esse contraste também se reflete na distribuição por seções, detalhada

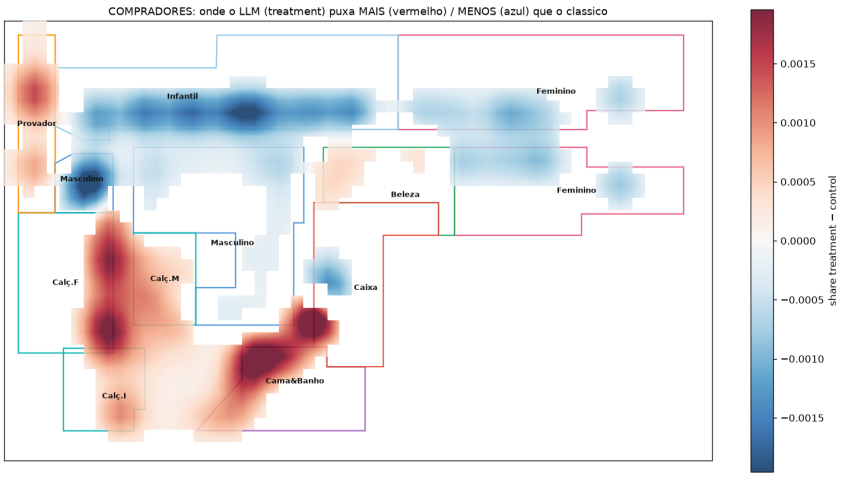
adiante. A Figura 10 mostra a diferença entre as abordagens generativa e clássica: em vermelho, as zonas mais circuladas pelos clientes do generativo; em azul, as mais circuladas no clássico.

Figura 9 – Mapa de calor de ocupação da loja pelos compradores: clássico (esquerda) e generativo (direita).



Fonte: Elaborada pelo autor.

Figura 10 – Diferença de ocupação entre as abordagens (generativo menos clássico), apenas compradores.



Fonte: Elaborada pelo autor.

Esse contraste pode ser quantificado separando, para a abordagem generativa, quanto o agente visita cada seção do quanto de fato compra nela (Tabela 22). Dois padrões emergem, e ambos apontam para a decisão de destino, e não para a conversão. O primeiro é o comportamento de vitrine em calçados: é a seção mais visitada (35,2% das idas), mas a de menor conversão por visita (5%), pois o agente circula e observa, comprometendo-se pouco, e só alcança 19,8% do valor porque o calçado, quando comprado, é um item caro. O segundo esclarece a lacuna do infantil: longe de ser uma falha de conversão, é a seção de maior conversão por visita (45%), embora seja a menos visitada (5,5%). O baixo mix do infantil nasce, portanto, de o agente

escolher ir até lá de forma rasa e solo, e não de deixar de comprar quando vai, o que reforça a limitação de dinâmica familiar discutida na Seção 7.3.2.

Tabela 22 – Padrão de visitas, conversão por visita e mix de valor por departamento (abordagem generativa).

Departamento	Visitas (%)	Conversão/visita	Mix valor (%)
Feminino	23,1	26%	36,9
Masculino	6,8	28%	14,7
Infantil	5,5	45%	12,6
Calçados	35,2	5%	19,8
Cameba	19,9	17%	13,2
Beleza	7,0	13%	3,0

Fonte: Elaborada pelo autor. A conversão por visita é a fração de idas à seção que resultaram em compra, distinta da conversão da loja por cliente.

O mesmo padrão, medido para a abordagem clássica (Tabela 23), evidencia por contraste o que a camada cognitiva acrescenta. Enquanto no generativo a conversão por visita varia de 5% em calçados a 45% no infantil, quase dez vezes, no clássico ela é praticamente plana, situando-se entre 10% e 19% em todos os departamentos, sem que nenhum se destaque. A razão é estrutural: o clássico navega em proporção à área de cada setor e decide a compra por sorteio dentro do orçamento, de modo que cada departamento se comporta como o mesmo evento aleatório, indiferente ao seu significado. Ele não distingue a vitrine da missão, não ”sabe” que o calçado é um setor de passagem e observação, nem que a roupa infantil, quando visitada, corresponde a uma intenção deliberada. Essa uniformidade também explica por que o clássico visita e vende no infantil em patamar próximo ao real (21,4% das idas e mix em torno de 18%): ele chega lá pela área, e não por propósito. A dispersão da conversão por visita no generativo, ao contrário, é a assinatura da intenção que a decisão contextual introduz, e é o que dá sentido à leitura espacial da jornada.

Tabela 23 – Padrão de visitas, conversão por visita e mix de valor por departamento (abordagem clássica).

Departamento	Visitas (%)	Conversão/visita	Mix valor (%)
Feminino	29,4	16%	30,4
Masculino	17,3	14%	21,5
Infantil	21,4	19%	17,8
Calçados	19,2	10%	21,1
Cameba	7,6	14%	7,2
Beleza	5,1	10%	2,0

Fonte: Elaborada pelo autor. Visitas e conversão por visita instrumentadas em 8 *seeds* (coorte de pico), sob a mesma definição da Tabela 22.

Em síntese, a camada generativa produz uma dinâmica espacial mais rica e exploratória, mas o seu critério mais seletivo de *onde ir* e *o que comprar* redistribui a atividade entre os departamentos, afastando-a do padrão clássico calibrado ao mix real.

7.3.5 Análise qualitativa das decisões

O transcript, módulo de coleta de dados citado na Seção 7.2, permite a leitura da jornada do cliente: motivo da ida à loja, navegações, compras e conversas, o que possibilita identificar as incongruências nas decisões tomadas pelos clientes.

Em versões anteriores da simulação verificou-se a compra de itens do sexo oposto por clientes. A fim de esclarecer essa anomalia na lógica de compra, introduziu-se a necessidade de uma justificativa na compra do item e da indicação de para quem ele se destina. Verificou-se que não se tratava de um *bug*, mas sim de presentes: pessoas comprando para outras pessoas, evento que ocorre em cerca de 19% das compras, apesar da ausência de uma dinâmica familiar rica.

Por outro lado, um problema real capturado pelo transcript foi a invenção de produtos e preços que não faziam parte do catálogo original. A primeira medida para resolver esse problema foi o desenvolvimento de uma trava mecânica que, por meio de um filtro regex, fazia o cliente levar o item do mesmo setor que possuísse o nome mais próximo ao oferecido e que coubesse no seu orçamento, como no exemplo da Figura 11:

Figura 11 – Exemplo de alucinação do vendedor: o item e o preço oferecidos não constavam no catálogo.

Vendedor: Tenho um tênis casual, o “Tênis Comfort”, por R$99,99.
Cliente: Estou interessado. Pode me trazer o número 42?
→ *Registrado:* `TENIS CANO ALTO`*, R$191,57 (outra peça, quase o dobro do valor oferecido).*

Fonte: Elaborada pelo autor.

Porém, isso adicionou outro problema: o preço que o cliente concordava em pagar e o do item que ele efetivamente levava eram bem diferentes. Logo, a verdadeira solução foi o envio do catálogo em todos os *rounds* de conversa do vendedor. Enviá-lo somente no início fazia com que, se o cliente recusasse a primeira oferta (que era real), as seguintes fossem alucinações. Com o catálogo em todos os turnos, o problema foi praticamente eliminado: apenas 0,3% das conversas ainda apresentam um preço inexistente.

No conjunto, o comportamento gerado é majoritariamente plausível: a cadeia história, navegação e compra é coerente, As inconsistências presentes na conversa foram corrigidas ou contidas pelo filtro. Sobretudo, decisões que, isoladas, parecem falhas

revelam-se plausíveis na jornada completa, e é essa leitura de conjunto que sustenta a avaliação da camada cognitiva. Para ilustrar, apresentam-se três jornadas completas e distintas entre si, da persona à saída da loja.

7.3.5.1 Letícia Vargas

Persona: 32 anos, solteira, sem filhos, escolaridade fundamental, parda, casa própria, trabalha fora, orçamento de R$488,32.

> *"Oi, sou a Letícia e hoje tô aqui na Lojas ******* porque preciso de umas roupas novas pro trabalho. Eu costumo usar o que tenho, mas dei uma olhada no meu guarda-roupa e percebi que as camisetas já tão bem desgastadas e eu quero algo novo pra me sentir mais confiante. Além disso, já que estamos em abril, tá começando a esfriar um pouco, então, uma jaqueta leve ia cair bem. Também pensei que poderia dar uma olhada em algumas coisas pro meu quarto, talvez jogar umas novas toalhas ou roupas de cama, porque a gente merece um lar mais arrumado, né? Estou animada pra ver o que a loja tem!"*

Jornada: No feminino, leva uma `JAQUETA JEANS` e duas `BLUSA MANGA LONGA`, todas justificadas pelo trabalho e pela autoconfiança ("pra me sentir mais confiante"). No cama, mesa e banho, renova o quarto com uma `COLCHA CASAL` e um `LENÇOL AVULSO` ("a gente merece um lar mais arrumado"), e ainda aceita uma `TOALHA BANHO` oferecida pelo vendedor, item que ela mesma havia mencionado querer ver. Passa por calçado e beleza sem comprar, mantendo o foco no que veio buscar.

Desfecho: Uma cesta dividida entre o guarda-roupa de trabalho e o enxoval da própria casa, coerente com a jovem que mora sozinha e se permite pequenas melhorias, encerrando a visita com R$60,51 dos R$488,32 iniciais.

7.3.5.2 Marcelo Bortolini

Persona: 52 anos, casado, com filhos, escolaridade fundamental, branco, casa própria, trabalha fora, orçamento de R$210,50.

> *"Oi, sou o Marcelo. Hoje eu vim na Lojas ******* porque a gente estava precisando de algumas roupas para o outono, principalmente para os meus filhos. Com a chegada da nova estação, sempre bom dar uma renovada no armário, né? Eles cresceram rápido e já cabem em tamanhos maiores. Além disso, eu mesmo estou precisando de algumas camisetas e um novo par de tênis, que os meus estão bem usados. Acho que também vou dar uma olhada nos itens de cama-mesa-banho, porque a gente está querendo trocar alguns jogos de cama. A esposa sempre reclama que os jogos que temos estão bem desgastados, então hoje aproveito que estou aqui."*

Jornada: Começa pelo infantil, justamente pelos filhos: leva uma calça de moletom para o filho (“vai ser boa para o meu filho usar”), mas no calçado infantil rejeita o que vê, pois no seu orçamento só cabem itens de menina, quando procurava para os meninos. Em seguida leva um `JOGO LENÇOL CASAL` e um `LENÇOL AVULSO` para a casa (“a esposa vai gostar”) e um `KIT MEIA` para si (“meias sempre ficam em falta aqui em casa”), além de procurar um tênis para si repetidamente sem concluir. Apesar de ter vindo principalmente pelos filhos, destina a eles uma única peça e concentra o gasto em si e na casa.

Desfecho: Uma cesta voltada sobretudo à casa e a si mesmo, um retrato em miniatura da sub-representação do infantil: o agente adulto vem pela família, mas destina às crianças apenas uma fração do que gasta. Encerra quase no limite do orçamento, com apenas R$10,93 dos R$210,50 iniciais.

7.3.5.3 Sandra Dalla

Persona: 62 anos, casada, com filhos, escolaridade médio, parda, casa própria, orçamento de R$239,32.

> *“Oi, meu nome é Sandra Dalla e estou aqui na Lojas ******* porque preciso renovar um pouco meu guarda-roupa para a temporada de inverno que já está chegando. Além disso, tem um aniversário de uma das minhas netas se aproximando e eu sempre gosto de levar um presente legal. Estou aqui porque também quero encontrar algo confortável, mas bonito para mim, já que sempre me preocupo em me sentir bem e me vestir adequadamente, mesmo com a rotina corrida do dia a dia. A loja tem ótimas opções a preços acessíveis e, como sou cliente frequente, sempre tenho uma certa confiança nas roupas que encontro aqui. Espero achar uma blusa quentinha e talvez um sapato bonito para combinar!”*

Jornada: No feminino, leva uma `BLUSA MANGA LONGA` e, no calçado feminino, uma `SAPATILHA`, ambas justificadas como para o próprio guarda-roupa (“combina bem com as minhas roupas”). Em seguida procura o presente da neta no calçado e no vestuário infantil, mas não se decide (“não achei nada que eu quisesse levar para a minha neta”); passa por beleza e por cama, mesa e banho sem comprar e segue ao caixa. A intenção de presentear atravessa toda a visita, mas não se concretiza.

Desfecho: Leva duas peças para si e encerra a visita com R$89,00 dos R$239,32 iniciais, sem o presente da neta.

7.4 Experimentos contrafactuais

Esta seção apresenta os resultados dos experimentos contrafactuais, nos quais se avalia a plausibilidade do comportamento do modelo diante de uma configuração de

loja que não existe na realidade. É a resposta à pergunta “e se eu aplicar determinada técnica, o que acontece?”, que caracteriza o uso do simulador como um *sandbox* de apoio à decisão, no qual o gestor testa uma estratégia sem risco operacional.

São avaliadas três estratégias: mudança de layout, promoção de itens e diferentes tons no atendimento. Cada uma segue o mesmo protocolo A/B, no qual se compara a execução de três *seeds* com a técnica aplicada à execução com a configuração padrão da loja. Essas *seeds* (1, 7 e 42) foram escolhidas aleatoriamente e mantidas fixas em todos os contrafactuais, reaproveitando valores já empregados no experimento principal, o que assegura a reprodutibilidade e a comparabilidade entre os cenários; o número reduzido de execuções decorre do custo computacional da abordagem generativa, discutido na seção de limitações. A análise não tem a ver com precisão numérica, mas sim com a direção e a plausibilidade dos resultados ocasionados pela intervenção. Nesses experimentos, a abordagem clássica fica em segundo plano, uma vez que o resultado depende mais do contexto do que de regras codificadas mecanicamente: a camada generativa enxerga a intervenção e ajusta a decisão de forma emergente, enquanto o clássico só reage àquilo que foi explicitamente programado.

7.4.1 Layout

O primeiro cenário contrafactual é o de alteração de layout. Nele, o agente generativo passa a perceber o ambiente em que se encontra, recebendo, na decisão de navegação, a lotação de cada setor (vazio, tranquilo, movimentado ou lotado) e o total de pessoas na loja. A lotação é a razão entre quantas pessoas estão no setor e a sua capacidade, definida como a área de venda do departamento dividida por 2,5 m^2 por pessoa, valor que reflete a norma de densidade confortável (cerca de uma pessoa a cada 2 a 3 m^2 de área de venda); como a capacidade depende do tamanho do departamento, um setor maior comporta mais gente antes de parecer cheio, o que introduz, de forma indireta, a percepção de espaço. A Tabela 24 mostra os limiares que traduzem essa razão em rótulo qualitativo, na forma como uma pessoa percebe um ambiente.

Tabela 24 – Faixas de ocupação e o rótulo qualitativo de lotação percebido pelo agente.

Rótulo de lotação	**Ocupação (pessoas ÷ capacidade)**
Vazio	menor que 25%
Tranquilo	de 25% a 60%
Movimentado	de 60% a 90%
Lotado	90% ou mais

Fonte: Elaborada pelo autor.

Somente ao ligar o efeito de percepção de fluxo, não houve grande diferença perceptível para o *baseline*: com as áreas normais, quase nenhuma navegação chega a citar a lotação. Através disso, percebeu-se que as áreas simuladas eram grandes de-

mais para chegarem a passar uma ideia de fluxo alto e praticamente todos os setores eram lidos como vazios. Optou-se por reduzir a área de todos os departamentos a um décimo, com exceção do departamento infantil, que, em relação aos demais, pareceria "vazio". Com essa decisão, o efeito se tornou bem mais perceptível nas justificativas de navegação, como mostrado na Tabela 25 e na Figura 12.

Tabela 25 – Percentual de decisões de navegação que citam a lotação/fluxo na justificativa.

Cenário	Navegações	Citam fluxo	%
Áreas normais (seed 1)	516	17	3,3
Áreas normais (seed 7)	505	20	4,0
Áreas normais (seed 42)	514	22	4,3
Infantil ampliado (seed 1)	439	107	24,4
Infantil ampliado (seed 7)	469	110	23,5
Infantil ampliado (seed 42)	534	119	22,3

Fonte: Elaborada pelo autor.

Figura 12 – Justificativas de navegação relacionadas ao fluxo de pessoas, com a percepção de layout ligada.

→ **Infantil:** "Preciso renovar as roupas das crianças e esse setor está vazio, ideal para conseguir olhar com calma."

→ **Infantil:** "Estou precisando de roupas novas para os meus filhos e este setor está tranquilo, posso olhar com calma."

→ **Beleza:** "Quero aproveitar que o setor de beleza está vazio e ver se encontro algum acessório ou maquiagem que eu possa usar no dia a dia."

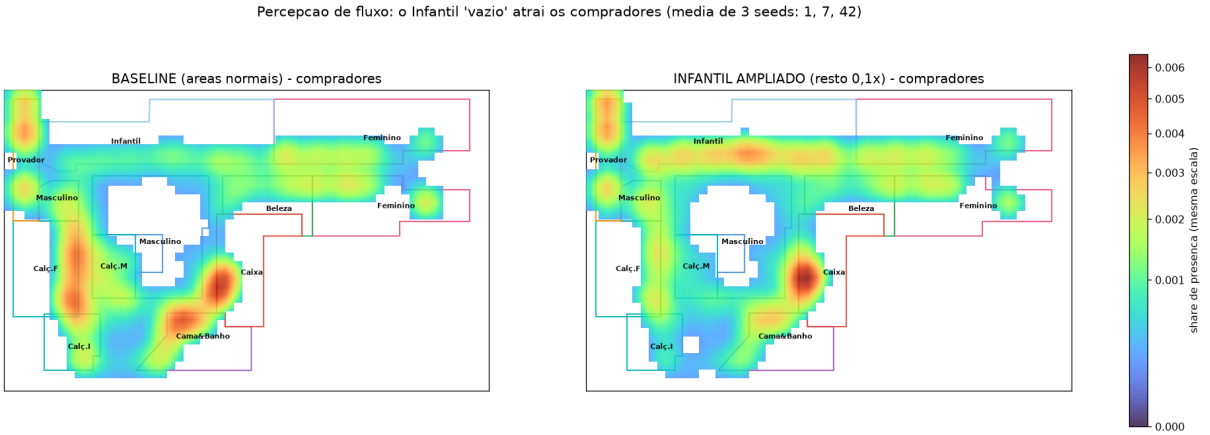
→ **Feminino:** "Quero ver roupas novas para renovar meu guarda-roupa e o setor feminino está movimentado, então tem mais opções."

Fonte: Elaborada pelo autor.

A grande maioria dos clientes preferiu ir aos setores mais tranquilos, porém é possível ver, no último caso, algo que foge do padrão, mas que não é incongruente: um cliente associou a grande movimentação do setor feminino como justificativa para ter mais variedade de roupas.

Como mostra a Figura 13, que agrega as três sementes, o fluxo de compradores se desloca visivelmente para o Infantil, agora o setor mais espaçoso. Esse deslocamento se traduz no mix: em média, a participação do Infantil sobe de 15,4% para 23,7%, aproximando-se do valor real (23,3%). O efeito é consistente na direção nas três sementes, embora varie em magnitude. Como a conversão e o ticket permanecem estáveis, trata-se de uma redistribuição, e não de mais ou menos compra.

Figura 13 – Mapa de calor de ocupação pelos compradores, média das *seeds* 1, 7 e 42: *baseline* (esquerda) e com o Infantil ampliado (direita).



Fonte: Elaborada pelo autor.

Para separar os efeitos cognitivo e mecânico dos resultados, aplicou-se o mesmo redimensionamento de áreas à abordagem clássica (Infantil continuou com área normal, e as demais reduzidas a 1/10). No clássico, o efeito foi na mesma direção esperada que o generativo (Tabela 26), porém de forma extrema, quase proporcional à sua mudança de área, reação esperada devido ao seu sorteio de navegação ponderado. A participação do Infantil no mix saltou de 17,8% para 73,0%. Na loja real, os clientes não concentrariam suas compras em somente um setor, apenas devido ao seu tamanho; quanto ao *baseline*, a ponderação por área provavelmente faz sentido, porque os departamentos da loja são divididos com base no seu público, e não como resultado dele.

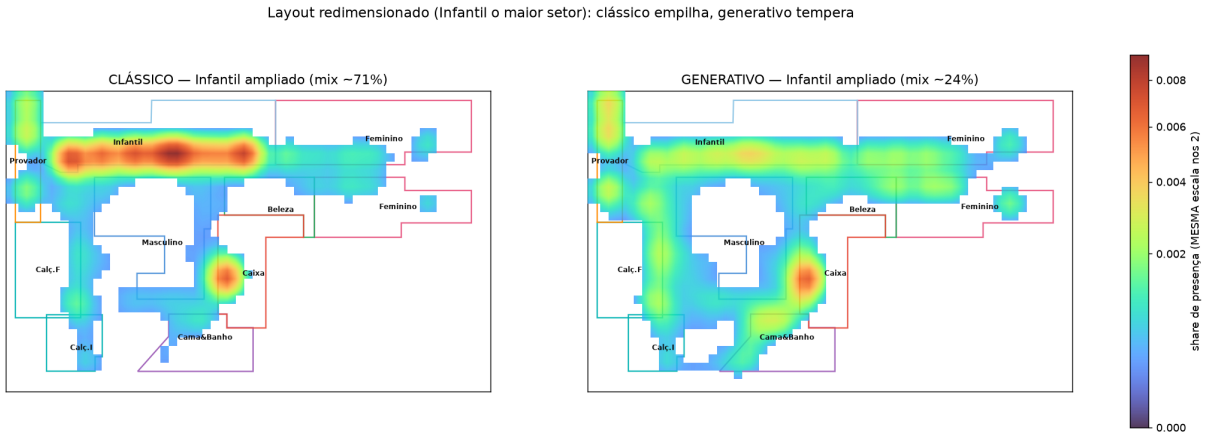
Tabela 26 – Efeito do redimensionamento (Infantil ampliado) sobre a participação do Infantil no mix, por abordagem

Abordagem	***Baseline***	**Layout**	Δ
Clássico (8 *seeds*)	17,8%	73,0%	+55,2 p.p.
Generativo (3 *seeds*)	15,4%	23,7%	+8,3 p.p.

Fonte: Elaborada pelo autor.

A comparação através da Figura 14 evidencia que a camada cognitiva tempera o "puxão geométrico" com comportamento, produzindo uma resposta mais plausível ao *layout* do que a proporcionalidade cega do clássico.

Figura 14 – Mapa de calor das duas abordagens sob o Infantil ampliado



Fonte: Elaborada pelo autor.

7.4.2 Promoção

O segundo cenário contrafactual permite definir uma promoção em determinado departamento ou item específico, com a possibilidade de estipular um desconto de 0 a 80%, em incrementos de 5%. Esse cenário afeta a percepção dos clientes ao receberem o catálogo de itens do departamento por onde navegaram, indicando se o produto está em promoção, qual o percentual de desconto e os seus valores antigo e atual. Os vendedores também podem utilizar a promoção a seu favor, uma vez que podem citá-la em suas conversas para tentar persuadir o cliente.

Como visualizado no experimento principal, grande parte das visitas era em direção ao departamento de calçados, mas, por ser um departamento de produtos com preços altos, a conversão era muito rara. Decidiu-se que esse seria o setor-alvo da promoção, e foi aplicado um desconto de 20% em todos os calçados: femininos, masculinos e infantis.

O efeito da promoção foi o aumento da compra de calçados, resultando numa maior representação desse setor no mix de vendas, como apresentado na Tabela 27. Analisando o transcript, é possível visualizar a diferença no comportamento dos vendedores, que usaram os preços promocionais para persuadir os clientes a comprar, e os clientes, mesmo sozinhos, foram convencidos a gastar parte de seus orçamentos numa peça um pouco mais cara (Figura 15).

Figura 15 – Exemplo de conversa sobre a promoção nos calçados.

Vendedor: Oi, tudo bem? Vi que você pegou uma blusa e um lençol. Posso ajudar com mais alguma coisa?
Cliente: Oi, tudo certo. Estou só olhando os calçados agora. Eles têm alguma promoção?
Vendedor: Sim, temos várias promoções nos calçados! Por exemplo, temos o Tênis Lifestyle Premium Feminino, que está de R$252,39 por R$201,91, e o Tênis Cano Alto, que virou de R$191,57 por R$153,26. Posso mostrar mais opções se você quiser!
Cliente: Estou interessada no Tênis Cano Alto, pode me trazer um número 37?

Fonte: Elaborada pelo autor.

Quanto à abordagem clássica, o mesmo desconto foi aplicado nesse setor; porém, como a compra é definida por sorteio, a mesma quantidade de calçados continuou sendo adquirida, agora a preços menores, ocasionando o efeito contrário ao esperado, isto é, uma leve queda no valor em vez de aumento, ainda assim dominada pelo ruído entre as *seeds*.

Tabela 27 – Efeito da promoção sobre a participação do departamento no valor vendido, o ticket médio e a cesta

Abordagem / *seed*	***Base.* (%)**	**Promo. (%)**	**Δ (p.p.)**	**Ticket (R$)**	**Cesta**
Generativo, *seed* 42	15,3	45,1	$+29{,}8$	$163 \rightarrow 179$	$2{,}78 \rightarrow 2{,}37$
Generativo, *seed* 1	19,4	39,7	$+20{,}3$	$155 \rightarrow 229$	$2{,}37 \rightarrow 3{,}08$
Generativo, *seed* 99	15,5	34,1	$+18{,}6$	$173 \rightarrow 227$	$2{,}86 \rightarrow 2{,}97$
Generativo (média, n=3)	**16,7**	**39,6**	$\mathbf{+22{,}9 \pm 4{,}9}$	$\mathbf{163 \rightarrow 212}$	$\mathbf{2{,}67 \rightarrow 2{,}81}$
Clássico (média, n=7)	**20,6**	**16,7**	$\mathbf{-3{,}9 \pm 10{,}1}$	$\mathbf{187 \rightarrow 182}$	$\mathbf{2{,}41 \rightarrow 2{,}36}$

Fonte: Elaborada pelo autor.

Um efeito desse contrafactual na abordagem generativa foi o aumento do ticket médio, que em duas das três *seeds* chegou a atingir cerca de R$227 e R$229, perto da média real da loja. Esse ganho, no entanto, vem sobretudo do valor das peças, e não do tamanho da cesta: enquanto o ticket médio subiu de R$163 para R$212, a cesta aumentou de forma estável (de 2,67 para 2,81 itens por cupom), chegando a cair em uma das *seeds*. Ou seja, o cliente não leva muitos itens a mais, e sim substitui parte da cesta por um calçado de maior valor, agora acessível pelo desconto. Isso evidencia que talvez o orçamento não fosse o bloqueador absoluto da representação da cauda rica, e que o efeito da compra impulsiva por promoções de fato afeta a representação dos indicadores.

7.4.3 Atendimento

O terceiro cenário contrafactual tem como objetivo visualizar a reação do agente cliente ao ser abordado com diferentes tons por parte do agente vendedor, uma reação ao contexto que só a abordagem generativa é capaz de produzir. Para isso, foram testados quatro tipos diferentes de abordagem, que constam na Tabela 28.

Tabela 28 – Texto injetado no *prompt* da LLM indicando o tom a ser usado.

Tom	**Texto injetado no *prompt***
Neutro (*baseline*)	"SEU JEITO DE ABORDAR é NEUTRO/INFORMATIVO: você responde o que perguntam e dá informação, sem empurrar nem insistir."
Incisivo	"SEU JEITO DE ABORDAR é INCISIVO/INSISTENTE: você pressiona, cria urgência e insiste para fechar a venda. Não desiste fácil."
Calmo	"SEU JEITO DE ABORDAR é CALMO/CONSULTIVO: você é atencioso, escuta, sugere sem pressionar e deixa o cliente à vontade."
Caloroso	"SEU JEITO DE ABORDAR é CALOROSO/ENTUSIASMADO: você elogia bastante, anima e cria conexão emocional."

Fonte: Elaborada pelo autor.

Por meio da análise das conversas no transcript, é possível verificar alguns padrões que se repetem nas abordagens de um mesmo tom. A seguir, apresenta-se um exemplo de conversa de cada tom, com o seu padrão explicado (Figuras 16 a 19):

Figura 16 – Exemplo do tom neutro: o vendedor informa as opções pedidas, sem elogiar nem pressionar.

Vendedor: Oi! Estou por aqui se precisar de ajuda para escolher alguma peça. O que você procura hoje?
Cliente: Oi, obrigado! Estou procurando algumas camisetas novas e talvez uma calça mais leve para o outono.
Vendedor: Legal! Temos algumas opções de camisetas, como a camiseta básica por R$64,98 e a camiseta manga curta adulta verão por R$82,97. Para calça leve, a calça esportiva adulta permanente a R$111,75 pode ser uma boa escolha. O que acha dessas?
Cliente: Oi, eu gostei da camiseta básica por R$64,98. Pode me trazer uma G, por favor?

Fonte: Elaborada pelo autor.

Figura 17 – Exemplo do tom incisivo: o vendedor insiste com urgência (escassez) sobre a mesma peça, e o cliente recusa repetidamente, sem fechar a compra.

Vendedor: E aí, tudo certo? Olha, dá uma olhada nessa camiseta manga curta verão por R$82,97! Pode ser uma boa, né?
Cliente: Parece interessante, mas acho que vou continuar dando uma olhada em outras opções primeiro.
Vendedor: A camiseta manga curta verão é sucesso garantido! O pessoal tá amando e é limitada, viu? Não dá pra perder essa, hein? Que tal dar uma chance agora?
Cliente: Obrigada, mas vou continuar olhando. Não quero me precipitar nessa escolha.
Vendedor: Sem problema, mas só avisando, essa camiseta é a mais procurada! Se você perder, pode ficar sem. Já pensou? Não demora, hein?
Cliente: Vou continuar olhando mais algumas camisetas antes de decidir.
Vendedor: Beleza, mas não esquece que a camiseta manga curta verão tá voando das prateleiras! Não quero que você fique sem, hein?
Cliente: Vou continuar dando uma olhada, obrigado!

Fonte: Elaborada pelo autor.

Figura 18 – Exemplo do Tom calmo: o vendedor observa a cesta e sugere, sem pressão, uma peça que combina.

Vendedor: Oi, tudo bem? Vi que você escolheu uma blusa e uma legging. Posso te mostrar alguns tênis que combinam bem com essas peças?
Cliente: Oi, tudo bem! Eu estava pensando em dar uma olhada em tênis. Que opções você tem?
Vendedor: Que bom que você está aqui! Tenho algumas opções legais. Olha, temos o TENIS INF MENINA por R$144,58 e o TENIS INF MENINO por R$196,60. Também temos o TENIS BEBE MENINA por R$129,71 e o TENIS BEBE MENINO por R$101,82. Algum desses te interessa?
Cliente: Estou interessada no TENIS INF MENINA, parece um bom complemento para a legging que escolhi.

Fonte: Elaborada pelo autor.

Figura 19 – Exemplo do tom caloroso: o vendedor elogia a escolha e cria conexão emocional.

Vendedor: Oi, tudo bem? Essa blusa de manga longa tá maravilhosa! Você já pensou em combinar com alguma peça legal? Olha, temos várias opções! Posso te mostrar?
Cliente: Oi! Estou sim, obrigada. Eu realmente gostei dessa blusa de manga longa. O que você sugere para combinar com ela?
Vendedor: Que bom que você gostou! Para complementar, eu sugiro uma colcha casal, fica lindo e aconchegante! Só R$68,45. O que acha?
Cliente: Uma colcha casal é uma ótima ideia! Não estava pensando em algo assim, mas gostei da proposta. Vou levar!

Fonte: Elaborada pelo autor.

Cabe frisar que o ruído pode ter corrompido parte da análise devido à baixa quantidade de execuções (3 *seeds*); ainda assim, os diferentes tons retornaram resultados coerentes, apresentados na Tabela 29. Comportamentos como sugerir combinações ou listar opções aparecem em proporção parecida em todos os tons, sendo próprios do vendedor; o que de fato distingue os tons, medido no transcript, são o elogio e a criação de urgência. O tom neutro, tomado como padrão, faz apenas o básico, sem elogiar nem pressionar, e produziu os menores KPIs dentre os quatro. O tom calmo, suave e consultivo, eleva os indicadores apenas de forma leve, mantendo-se próximo do neutro. Já os tons que mais convenceram os clientes a levarem itens foram o incisivo e o caloroso. O incisivo é o que mais cria urgência (presente em 24% de suas conversas, contra 2% do neutro) e produziu o maior ticket e a maior cesta; esse tom apareceu com diferentes níveis de insistência, e a mais intensa nem sempre foi bem recebida: no exemplo de conversa citado, o próprio cliente registra a percepção do tom (“Me senti um pouco pressionado, mas estou decidido a olhar mais opções”) e, mesmo tendo orçamento para a camiseta, não se deixou levar. O caloroso, por sua vez, é o que mais elogia (78% das conversas, contra 11% do neutro) e converteu quase tanto, porém por meio da conexão emocional.

Tabela 29 – Efeito do tom de abordagem do vendedor sobre os indicadores.

Tom	Conversão (%)	Ticket (R$)	Itens/cupom	Persuasão (%)
Neutro (*baseline*)	29,3 $\pm$ 2,4	159,0 $\pm$ 3,3	2,5 $\pm$ 0,2	10,1 $\pm$ 1,1
Incisivo	31,3 $\pm$ 2,5	197,2 $\pm$ 12,3	3,0 $\pm$ 0,2	15,0 $\pm$ 1,2
Calmo	30,7 $\pm$ 0,8	180,8 $\pm$ 18,9	2,8 $\pm$ 0,3	12,2 $\pm$ 1,5
Caloroso	30,5 $\pm$ 2,9	181,9 $\pm$ 19,8	2,6 $\pm$ 0,2	14,6 $\pm$ 2,8

Fonte: Elaborada pelo autor. Persuasão = fração das conversas que terminaram em compra.

7.5 Discussão dos resultados

A seção atual reúne os resultados apresentados nas seções anteriores e analisa o que significam em conjunto. O quadro que emerge é o seguinte: em determinadas análises as abordagens produzem resultados parecidos, em outras o generativo se sobressai e em outras o clássico. Logo, é impossível indicar uma abordagem para todas as análises.

Começando pelas análises em que ambas as abordagens empatam, temos os indicadores presentes na Tabela 17: a conversão, cujo valor é ancorado nos da loja real; o ticket, em que ambas geram valores muito parecidos para indicar um vencedor; a cesta, em que, apesar de a generativa ter chegado mais próxima da real, o clássico consegue atingir esse valor também com calibração; e o faturamento, em que o clássico chega mais próximo, ainda que ligeiramente abaixo do real. Em caso de preciosismo, pode-se indicar que o clássico mimetiza melhor esses valores, mas, como a vantagem não é expressiva e levando em conta apenas os resultados (e não o custo, no qual o clássico leva vantagem por não depender de API), declara-se um empate, com um bom espaço de melhora em ambas as abordagens, provavelmente ocasionado pela má representação do orçamento de compra dos clientes.

O ranking de itens (Tabela 20) é a métrica que ambas as abordagens representam com pior fidelidade, cada uma à sua maneira, mas ambas atingindo um resultado não satisfatório; a abordagem clássica não tem como aumentar o seu poder de representação, enquanto a generativa pode recorrer a métodos para representar melhor o cliente. Quanto à representação do mix de departamento (Tabela 19), a abordagem clássica tem grande vantagem: apesar de sub-representar o departamento infantil, o sorteio ponderado pelo tamanho da zona representa muito bem o mix de forma geral no cenário base, com um erro cerca de 2,4 vezes menor que o produzido pela abordagem generativa. Esse alto erro generativo deve-se principalmente à concentração no departamento Feminino e cama, mesa e banho. No experimento principal, portanto, a abordagem generativa representou os dados de forma semelhante ou menos fiel que a clássica, devido ao seu viés de saliência.

Nos experimentos contrafactuais de promoção, atendimento e layout, a abordagem generativa provou-se uma ferramenta útil para a análise contextual. Diante da promoção, os clientes apresentaram o comportamento de impulso, elevando a participação de calçados em cerca de 23 pontos percentuais (um aumento relativo de 137%) e o ticket para perto da média real (Tabela 27); diante do atendimento, os tons ativos elevaram o ticket e a persuasão, com assinaturas mensuráveis (o caloroso elogia em 78% das conversas contra 11% do neutro; o incisivo cria urgência em 24% contra 2%) (Tabela 29); diante do layout, ao perceber a lotação de cada setor, o generativo deslocou a sua circulação para o departamento mais espaçoso, elevando a participação do

departamento Infantil no mix de 15,4% para 23,7% e passando a citar explicitamente o fluxo em cerca de 23% das navegações, sem alterar a conversão nem o ticket, ou seja, uma redistribuição do fluxo (Tabela 25). O clássico, por reagir apenas ao que foi codificado mecanicamente, permaneceu inerte ao tom do vendedor; na promoção, a leve baixa nos valores dos calçados, além de ter sido diluída pelo ruído, ocorreu em direção contrária à esperada; no layout, a navegação clássica, proporcional à área, redistribui-se de forma mecânica e exagerada (o Infantil ampliado chega a concentrar 73% do mix), enquanto o generativo, por “perceber” a lotação e ponderá-la com a persona, desloca-se de forma moderada e plausível (para cerca de 24%).

Retomando a pergunta que motiva o trabalho, se a camada cognitiva representa melhor o consumidor do varejo, a resposta depende de qual análise está sendo realizada: se for uma análise dependente de contexto, a resposta é sim; se for uma análise possível de representar por meio de mecânicas e regras, a resposta é não, pois a abordagem clássica a reproduz com fidelidade equivalente, de forma mais rápida e sem custo de API. É essa constatação, e não a escolha de uma abordagem sobre a outra, que fundamenta o uso combinado das duas como ferramenta de *sandbox* de apoio à decisão.

8 CONCLUSÃO

Este trabalho utilizou dados de uma loja de varejo real para comparar a capacidade de representação do comportamento do consumidor, pondo à prova duas abordagens de simulação multiagente, a clássica e a generativa, a fim de verificar qual delas é mais adequada à produção de uma ferramenta de *sandbox* de apoio à decisão.

No experimento principal, que trata unicamente da representação dos números da loja, ambas as abordagens apresentaram valores muito próximos entre si e das métricas reais. Houve dois achados principais nesse experimento. O primeiro é que o clássico, apesar de não conseguir utilizar o contexto na tomada de decisão, apoiou-se na métrica correta: a navegação ponderada pelo tamanho da zona garantiu que ele chegasse a números próximos aos da abordagem generativa, sem custo de API e com execução extremamente rápida; além disso, a cesta, indicador em que o clássico ficou mais distante, poderia ter sido calibrada para se aproximar do comportamento esperado. O segundo diz respeito à abordagem generativa, que atingiu os mesmos valores que o clássico, mas de maneira diferente. Ambas as abordagens não representaram bem os produtos efetivamente comprados na loja: o clássico, por se distribuir de forma equilibrada entre todos os produtos, com exceção dos mais baratos; e o generativo, pela falta de uma dinâmica familiar e social rica, limitação que poderia ser mitigada através de engenharia de *prompt* e da inclusão de compra em grupo, a fim de enriquecer a persona e possivelmente redirecionar os clientes para os departamentos de forma correta.

Os três experimentos contrafactuais, layout, promoção e atendimento, foram realizados a fim de verificar a plausibilidade dos resultados produzidos pela abordagem generativa. Diante de uma promoção, o agente responde com comportamento de impulso, elevando a participação do setor e o ticket; diante de diferentes tons de atendimento, sua decisão muda de forma coerente com o estímulo; e diante da percepção de fluxo, desloca a sua circulação para os setores mais espaçosos, redistribuindo o mix sem alterar a conversão nem o ticket. O clássico, por reagir apenas ao que foi codificado mecanicamente, permanece inerte ou age de forma desproporcional a esses estímulos.

Do conjunto dos experimentos, verifica-se que a ferramenta de apoio à decisão ideal seria uma abordagem híbrida: em testes de natureza mecânica, como a variação do volume de entrantes ou a estimativa do mix por departamento no cenário base, a abordagem clássica seria priorizada, por não gerar custos de API e ser extremamente rápida; já em situações dependentes de contexto, como promoções, diferentes tons de atendimento e a reação a mudanças de *layout* pela percepção do fluxo de clientes, a abordagem generativa seria a escolhida, por captar a reação do consumidor que o clássico não reproduz ou reproduz de forma exagerada. Dessa forma, cada motor é empregado na dimensão em que se mostrou mais fiel, unindo a eficiência e o baixo custo do clássico à sensibilidade contextual do generativo em uma única ferramenta de *sandbox*.

8.1 Limitações

Os resultados de ambos os experimentos devem ser lidos sob a ótica das seguintes limitações do modelo:

- **Ausência de uma dinâmica familiar e social rica e da compra em grupo:** cada agente é simulado de forma isolada, e nenhum cliente tem ligações sociais ou de parentesco com os demais. Segundo o gerente, a compra em grupo é um evento recorrente na loja representada, e a sua ausência pode ser responsável pela sub-representação de determinados departamentos.
- **Baixo número de execuções nos contrafactuais:** cada uma das diferentes configurações da loja foi executada três vezes e, quanto menor esse número, mais a análise fica suscetível ao ruído entre *seeds*, o que pode gerar conclusões equivocadas.
- **Execuções restritas a horário de pico:** nos experimentos foram utilizadas apenas execuções de pico (14h – 15h), o ideal seria que as experimentações tivessem sido executadas durante um dia inteiro de loja, respeitando a base em que as medidas reais foram tomadas.
- **Fluxo do vendedor generativo de forma mecânica:** o vendedor generativo possui um papel de coadjuvante na simulação atual, sem poder decidir quem e quando abordar, decisão que permanece baseada em regras.
- **Modelos de LLM:** o presente trabalho não experimentou diferentes modelos de LLM, o que impede afirmar que os resultados não dependem do modelo utilizado.
- **Sensibilidade à engenharia de *prompt*:** o modelo passou por diversas versões até chegar à versão final e, ao longo delas, percebeu-se que as decisões da

LLM variavam intensamente conforme a quantidade e a qualidade da informação enviada, o que indica a necessidade de enriquecer a persona.

- **Métrica lexical para os padrões de atendimento:** as frequências de elogio e urgência por tom foram estimadas por correspondência de palavras-chave nas falas do vendedor, um *proxy*, e não uma classificação semântica validada.
- **Catálogo agregado por família, sem as variações de produto:** a complexidade do catálogo foi abstraída através da representação das 209 famílias dentro do Pareto, porém representa-se uma grande quantidade de itens por um preço médio representativo da categoria, fazendo com que um item de uma família que provavelmente caberia no orçamento de um cliente não caiba, impossibilitando a compra.
- **Memória limitada à visita atual:** o modelo garante a memória do cliente apenas durante a visita atual; o dado de recorrência da loja real é incluído de forma simplificada na persona, sem influenciar de maneira rica as suas ações.

8.2 Trabalhos futuros

As limitações acima apontam caminhos de melhoria para trabalhos futuros:

- **Modelar núcleos familiares e a compra em grupo:** marido e mulher, pais indo com os filhos, avós indo com os netos ou um grupo de amigos são alguns exemplos de possíveis grupos de compra a serem representados.
- **Ampliar o número de *seeds* por cenário:** como o ruído entre execuções decresce na proporção de $1/\sqrt{n}$, rodar cerca de 10 *seeds* por configuração reduz o erro padrão da média à metade do observado com 3, um equilíbrio razoável entre robustez estatística e o custo e a latência da abordagem generativa. Efeitos de pequena magnitude, como diferenças na conversão agregada, podem exigir ainda mais execuções, por permanecerem dentro da variabilidade entre *seeds*.
- **Simular o dia inteiro de loja:** executar os experimentos ao longo do dia útil completo, e não apenas no horário de pico, à medida que o custo dos modelos de linguagem diminua ou com o uso de modelos locais. Isso alinharia a simulação à base diária das métricas reais e permitiria observar a variação de perfil de consumidor e de volume entre as faixas horárias.
- **Vendedor com autonomia cognitiva:** dar à LLM a decisão de *quem* e *quando* abordar (priorizar clientes, ler sinais de interesse e gerir o próprio tempo), em

vez de a abordagem ser disparada por regras. Assim, abre-se espaço para analisar diferentes comportamentos sociais, como a criação de conversas dentro da equipe de vendedores.

- **Reduzir o custo da camada generativa:** com grande poder computacional disponível, torna-se possível executar modelos de LLM mais poderosos de forma local, verificando se a qualidade dos resultados se mantém.
- **Enriquecer as personas e refinar o *prompt*** para mitigar o viés de saliência e melhorar a fidelidade do mix, aproximando a distribuição emergente da real.
- **Validar as métricas qualitativas:** buscar maneiras mais rigorosas de validar as métricas qualitativas; o presente trabalho fez uso de um *proxy* lexical, baseado em palavras-chave.
- **Explorar diferentes versões do cálculo do orçamento de visita:** o cálculo atual não reproduz de forma satisfatória a cauda de tickets muito altos presente na distribuição dos valores de compra da loja real.
- **Representar os produtos no nível de SKU:** modelar o catálogo da loja de forma específica, representando todos os tipos de variação de produtos, cores e modelos, evitando comprimir todos os itens de uma família a um único preço médio, que, como evidenciado, acaba excluindo o cliente de uma família inteira por causa do valor representativo.
- **Memória persistente entre visitas:** modelar a memória do cliente ao longo de múltiplas visitas, e não apenas dentro de uma, aproveitando a rota *rounds* da API, já projetada para isso, permitindo analisar como a experiência acumulada influencia as visitas seguintes.

REFERÊNCIAS

Agno. **Agno**: Build AI Agents with Memory, Knowledge and Tools. New York: GitHub, 2024. Repositório online. Acesso em: 21 fev. 2026, https://github.com/agno-agi/agno.

ALVES, B. C.; BORN, M. B.; AGUIAR, M. S. d. Integrating Large Language Models as Cognitive Agents into the GAMA Platform for Urban Mobility Simulation. In: INTERNATIONAL FLAIRS CONFERENCE (FLAIRS-39), 39., 2026, Gainesville, FL, USA. **Proceedings...** Florida Online Journals: University of Florida, 2026. v.39.

BANKS, J.; CARSON, J. S.; NELSON, B. L.; NICOL, D. M. **Discrete-Event System Simulation**. 5th.ed. Harlow, England: Pearson, 2010.

BONABEAU, E. Agent-based modeling: methods and techniques for simulating human systems. **Proceedings of the National Academy of Sciences**, Washington, D.C., v.99, n.suppl 3, p.7280–7287, 2002.

BUGHIN, J.; HAZAN, E.; LABAYE, E.; MANYIKA, J.; DAHLSTRÖM, P.; RAMASWAMY, S. Why digital strategies fail. **McKinsey Quarterly**, New York, 2019.

CNC. **Estudo sobre a volatilidade no comércio varejista brasileiro**. Rio de Janeiro: Confederação Nacional do Comércio de Bens, Serviços e Turismo, 2024. Acesso em: 22 jan. 2026. Disponível em: <https://portaldocomercio.org.br>.

CURHAN, R. C. The relationship between shelf space and unit sales in supermarkets. **Journal of Marketing Research**, Thousand Oaks, CA, v.9, n.4, p.406–412, 1972.

Deloitte. **Deloitte Global Retail Outlook 2024**: Navigating Challenges and Embracing Opportunities. Londres: Deloitte Touche Tohmatsu Limited, 2024. Disponível em: https://www.deloitte.com/global/en/industries/consumer/analysis/global-retail-outlook.html. Acesso em: 22 fev. 2026.

DOS SANTOS, C. G. M.; SEABRA, I. C. N. ANÁLISE DAS DEMONSTRAÇÕES FINANCEIRAS: AVALIAÇÃO DE DESEMPENHO FINANCEIRO DE TRÊS EMPRESAS

DO SETOR VAREJISTA. **Amazon Business Research (ABR)**, Manaus, n.03, p.72–91, 2025.

eMarketer. **Global Retail E-commerce Forecast 2024**. New York: Insider Intelligence, 2024.

FARESE, L. S.; KIMBRELL, G.; WOLOSZYK, C. A. **Marketing Essentials**. Columbus, OH: Glencoe/McGraw-Hill, 2012.

HORIKOMI, T.; MIZUNO, T. Interaction-aware agent-based simulation of customer trajectories in retail stores with transformer architectures. **Scientific Reports**, London, v.15, n.1, p.38972, 2025.

IBGE. **Síntese de indicadores sociais**: uma análise das condições de vida da população brasileira 2023. Rio de Janeiro: IBGE, 2023. Cap. “Padrão de vida e distribuição de rendimentos”; PNAD Contínua 2022. Tabela 3 (p. 63).

IBGE. **Pesquisa Anual de Comércio (PAC) 2023**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2025. Acesso em: 22 jan. 2026. Disponível em: <https://www.ibge.gov.br>.

IBGE. **Sistema IBGE de Recuperação Automática — SIDRA**: Tabela 1737, IPCA (número-índice). Correção jun./2022 a abr./2026 (fator 1,177). Acesso em: 22 jan. 2026, `https://sidra.ibge.gov.br/tabela/1737`.

KATO, A.; OSHIMA, T.; NOMURA, T.; SUGAWARA, T. Constructing LLM-based Agent Model and its Application to Multi-Agent Simulation. In: THE 38TH ANNUAL CONFERENCE OF THE JAPANESE SOCIETY FOR ARTIFICIAL INTELLIGENCE (JSAI 2024), 2024, Hamamatsu, Japan. **Anais...** The Japanese Society for Artificial Intelligence (JSAI), 2024.

KELLER, K. L. Consumer-based brand equity. **Journal of Marketing**, Chicago, v.84, n.1, p.1–22, 2020.

KOTLER, P.; KARTAJAYA, H.; SETIAWAN, I. **Marketing 5.0**: Technology for Humanity. Hoboken: Wiley, 2021.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, London, v.521, n.7553, p.436–444, 2015.

LEMON, K. N.; VERHOEF, P. C. Understanding customer experience throughout the customer journey. **Journal of Marketing**, Chicago, v.80, n.6, p.69–96, 2016.

LI, B.; WEI, Q.; WANG, X. **Predicting Behaviors with Large Language Model (LLM)-Powered Digital Twins of Consumers**. Cambridge, MA: Marketing Science Institute, 2025. (25-135).

LIMA, L. M. d.; ARAÚJO, V. d. C.; SILVA, A. L. P. d. Padrões e evolução da demanda por vestuário no Brasil: uma análise a partir da POF entre 2008 a 2018. **International Journal of Business and Marketing (IJBMKT)**, São Paulo, v.9, n.2, p.37–57, 2024. QUAIDS sobre a POF-IBGE; recorte Sul.

LIU, P.; YUAN, W.; FU, J.; JIANG, Z.; HAYASHI, H.; NEUBIG, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. **ACM Computing Surveys**, New York, v.55, n.9, p.1–35, 2023.

MANSOUR, S.; PERELLI, L.; MAINETTI, L.; DAVIDSON, G.; D'AMATO, S. PAARS: Persona Aligned Agentic Retail Shoppers. In: WORKSHOP FOR RESEARCH ON AGENT LANGUAGE MODELS (REALM 2025), 1., 2025, Vienna, Austria. **Proceedings...** Association for Computational Linguistics, 2025. p.143–159.

MCCARTHY, J. **What is Artificial Intelligence?** Stanford, CA: Stanford University, 2007.

Mindtree. **Impact of Sales Associates on the Shopper's Purchase Journey**. Bengaluru: Mindtree; pesquisa por Grail Research, 2016. Compradores que interagem com um vendedor são 43% mais propensos a comprar e gastam 81% a mais.

MOE, W. W. Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream. **Journal of Consumer Psychology**, Mahwah, NJ, v.13, n.1–2, p.29–39, 2003.

MOHAMAD ALI, M. F.; ABUSTAN, M. S.; ABU TALIB, S. H. A Case Study of Malaysian Pedestrian Walking Speed at Shopping Malls in Kuala Lumpur, Malaysia using Human Behaviour Simulator (HBS). **International Journal of Integrated Engineering**, Batu Pahat, Malaysia, v.11, n.4, p.225–230, 2019.

NEWELL, A.; SIMON, H. A. Computer science as empirical inquiry: Symbols and search. **Communications of the ACM**, New York, v.19, n.3, p.113–126, 1976.

OLIVEIRA, G. B. L. L. d. **Desenvolvimento de um sistema multi-agente baseado em inteligências artificiais generativas para avaliação de qualidade em atendimento ao cliente**. 2025. 92 f.p. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) — Universidade Federal do Rio Grande do Norte, Natal, RN.

Opinion Box. **Consumo de Moda no Brasil 2024**. Infográfico de pesquisa (dez. 2024); 1.000 entrevistas, margem 3,1 p.p. Acesso em: jul. 2026, https://blog.opinionbox.com/wp-content/uploads/2024/12/OPB_infografico_moda_2024_final.pdf.

PANTANO, E.; PIZZI, G.; BILOTTA, E.; PANTANO, P. Enhancing store layout decision with agent-based simulations of consumers' density. **Expert Systems with Applications**, Amsterdam, v.182, p.115231, 2021.

PARENTE, J.; BARKI, E. **Varejo no Brasil**: gestão e estratégia. São Paulo: Atlas, 2014.

PARK, J. S.; O'BRIEN, J.; CAI, C. J.; MORRIS, M. R.; LIANG, P.; BERNSTEIN, M. S. Generative Agents: Interactive Simulacra of Human Behavior. In: ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY (UIST '23), 36., 2023, San Francisco, CA, USA. **Proceedings...** ACM, 2023. p.1–22.

RUMELHART, D. E.; MCCLELLAND, J. L. **Parallel Distributed Processing**: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press, 1986.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. 4th.ed. Boston: Pearson, 2021.

SBVC. **O Papel do Varejo na Economia Brasileira**. São Paulo: Sociedade Brasileira de Varejo e Consumo, 2024. Dados referentes ao ano base 2023. Disponível em: https://sbvc.com.br/estudo-11a-ed-o-papel-do-varejo-na-economia-brasileira-2024-sbvc/. Acesso em: 22 jan. 2026.

SHANKAR, V. How Artificial Intelligence (AI) is Reshaping Retailing. **Journal of Retailing**, Amsterdam, v.94, n.4, p.vi–xi, 2018.

SIEBERS, P.-O.; MACAL, C. M.; GARNETT, J.; BUXTON, D.; PIDD, M. **Discrete-Event Simulation is Dead, Long Live Agent-Based Simulation**. Lancaster, UK: AIM Research, Lancaster University Management School, 2008. (AIM Working Paper Series 023-2008).

SIMON, H. A. A Behavioral Model of Rational Choice. **The Quarterly Journal of Economics**, Oxford, v.69, n.1, p.99–118, 1955.

TAILLANDIER, P.; GAUDOU, B.; GRIGNARD, A.; HUYNH, Q.-N.; MARILLEAU, N.; CAILLOU, P.; PHILIPPON, D.; DROGOUL, A. Building, composing and experimenting complex spatial models with the GAMA platform. **GeoInformatica**, Cham, Switzerland, v.23, n.2, p.299–322, 2019.

UNDERHILL, P. **Why we buy**: The science of shopping–updated and revised for the Internet, the global consumer, and beyond. New York: Simon and Schuster, 2009.

V-Count. **Why Benchmarking Fitting Room Data is Important for the Retail Industry**. Permanência no provador ≈ 5 min 30. Acesso em: 2 jul. 2026, https://v-count.com/why-benchmarking-fitting-room-data-is-important-for-retail-industry/.

VAROTTO, L. F. História do varejo. **GV-executivo**, São Paulo, v.5, n.1, p.86–90, 2006.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2017, Long Beach, CA. **Anais...** Curran Associates, 2017. v.30.

VIAL, G. Understanding digital transformation. **MIS Quarterly**, Minneapolis, v.43, n.1, p.223–241, 2019.

WANG, G.; XIE, Y.; JIANG, Y.; MANDLEKAR, A.; XIAO, C.; ZHU, Y.; FAN, L.; ANANDKUMAR, A. **Voyager**: An Open-Ended Embodied Agent with Large Language Models. Disponível em: https://arxiv.org/abs/2305.16291. Acesso em: 24 jun. 2026, arXiv preprint arXiv:2305.16291.

WOOLDRIDGE, M. **An Introduction to MultiAgent Systems**. 2nd.ed. Chichester: John Wiley & Sons, 2009.

XI, Z. et al. **The Rise and Potential of Large Language Model Based Agents**: A Survey. Disponível em: https://arxiv.org/abs/2309.07864. Acesso em: 24 jun. 2026, arXiv preprint arXiv:2309.07864.

APÊNDICE A TEMPLATES DE PROMPT

Este apêndice reúne a engenharia de prompt realizada no projeto, apresentando cada um dos prompts enviados à LLM nos diferentes tipos de decisão: história, navegação, escolha de item e conversa (dos lados do cliente e do vendedor). Os valores entre chaves ($\{\ldots\}$) são preenchidos pelo GAMA em tempo de execução com os dados de cada agente.

A persona (Seção 6.8) é colocada no campo `description` do agente Agno; por ser um prefixo constante durante a visita, ela reduz o custo por meio do *prompt caching*. As seções seguintes mostram esse `description` e, em cada decisão, o prompt específico enviado.

A.1 Description Cliente

> Você é um CLIENTE que entrou numa loja popular de moda, em abril, no Rio Grande do Sul. Você NÃO é um assistente: você É essa pessoa. Você tem características próprias que formam a sua persona; aja de acordo com a sua personalidade. Ao longo da visita você decide PARA ONDE ir, O QUE comprar e INTERAGE com o vendedor, usando a sua memória da visita.
>
> VOCÊ É: {persona}. Você é {mulher/homem}[, cliente frequente da loja]. Hoje ({mês}) você está na {loja}. [Você veio porque: {história}]
>
> O que pesa em cada decisão EMERGE da sua persona, do seu motivo de hoje (a sua história), do que você já fez nesta visita (o que pegou, o quanto gastou, por onde passou) e do seu raciocínio; não há lista do que considerar nem roteiro de "se X então Y". Tudo o que você precisa está na sua situação atual; repare nela do seu jeito.

A.2 História

> Você ACABOU de chegar a esta loja. Pense em quem você é e na data de hoje e conte: por que VOCÊ veio a esta loja agora? Qual o seu

motivo principal? Escreva em “historia”, do seu jeito (1ª pessoa). Você ainda NÃO precisa decidir compras nem listar tudo; só o que de fato te trouxe aqui. A loja vende SÓ: roupas (feminino, masculino, infantil), calçados, cama-mesa-banho e maquiagem/acessórios.

Resposta: `{"historia": "..."}`

A.3 Navegação

O que há em cada setor: feminino = roupas femininas; masculino = roupas masculinas; infantil = roupas de bebê/criança/juvenil; calcados = calçados; cameba = cama, mesa e banho; beleza = maquiagem e acessórios.

Na chegada:

você ACABOU de chegar e vai escolher o PRIMEIRO setor.

Setores disponíveis: {lista de setores}. Pense e decida como ESSA pessoa: a partir de QUEM você é e do que já viu/fez, o que faz sentido agora? Não há resposta certa.

Em re-navegação:

você JÁ está circulando (agora no setor “{setor}”); passou por {já visitou}, já levou {cesta} e ainda pode gastar cerca de R$ {orçamento}; pode ir a outro setor, voltar a um que já viu, ou ENCERRAR (responda “sair”).

Setores disponíveis: {lista de setores}. Pense e decida como ESSA pessoa: a partir de QUEM você é e do que já viu/fez, o que faz sentido agora? Não há resposta certa.

Resposta: `{"ir_para": «setor ou 'sair'>", "motivo": "..."}`

A.4 Escolha de item

Você está OLHANDO uma arara/prateleira deste setor. Estas são as peças disponíveis que CABEM no que você pode gastar (você está vendo e examinando todas elas, nome e preço): {itens candidatos}. Você JÁ levou nesta visita: {cesta}. Seu estado nesta visita: {setor, orçamento, humor}.

Olhando essas peças, decida como sua persona decidiria: você LEVA (compra) UMA peça, ou NENHUMA. Antes de levar, avalie se você REALMENTE a compraria (do seu jeito). Seja coerente sobre PARA QUEM é a peça.

Resposta: `{"item": «peça ou 'nenhum'>", "motivo": "..."}`

A.5 Conversa (lado do cliente)

> SUA SITUAÇÃO AGORA: {setor, cesta, orçamento, humor, pensamentos de conversas}. {transcript até agora}
>
> Um vendedor se aproximou de você e disse: "{fala do vendedor}". Reaja como VOCÊ (1ª pessoa, curto) e tome sua decisão.

O bloco de decisão varia conforme o **orçamento**, verificado mecanicamente pelo GAMA (que lê o preço citado pelo vendedor) *antes* de montar o prompt. A LLM recebe apenas os caminhos que existem:

(a) A peça oferecida cabe no orçamento:

> O que o vendedor lhe ofereceu CABE no seu orçamento de hoje. Sua DECISÃO neste turno:
>
> "levar": você decide comprar a peça oferecida; seja coerente sobre para quem é a peça.
>
> - "continuar": você ainda quer conversar.
> - "terminar": você encerra a conversa e volta a olhar sozinho.

(b) A peça oferecida não cabe no orçamento:

> O que o vendedor lhe ofereceu NÃO cabe no seu orçamento de hoje: você não tem como levar. Sua DECISÃO neste turno:
>
> "continuar": você ainda quer conversar.
>
> - "terminar": você encerra a conversa e volta a olhar sozinho.

(c) Nenhum preço foi citado ainda (nada concreto a levar):

> Sua DECISÃO neste turno:
>
> "continuar": você ainda quer conversar.
>
> - "terminar": você encerra a conversa e volta a olhar sozinho.

Fecho comum aos três casos: "Caso queira seguir olhando sozinho, ou não tenha mais nada a tratar, escolha terminar. Não siga a conversa só por cortesia. Se for falar de uma peça, refira-se a ela pelo nome como o vendedor a chamou (não invente produtos); cor e tamanho você pode conversar naturalmente." Ao encerrar (terminar ou levar), o cliente também escreve em "pensamento" como se sentiu na conversa.

Resposta: `{"fala", "decisao": "continuar|terminar|levar", "item", "pensamento", "humor"}`

A.6 Description Vendedor

> Você é um VENDEDOR que trabalha numa loja popular de moda, em abril, no Rio Grande do Sul, e está atendendo um cliente. Você É

esse vendedor, com persona e estilo próprios. Converse em turnos, curto e natural, usando SÓ o que consegue OBSERVAR do cliente e o que ele disser (não invente o que ele não mostrou nem contou). Você conduz a conversa, mas NÃO decide se o cliente compra; ELE decide. Apresente e ofereça as peças na hora, pelo nome e preço; NÃO diga que vai buscar, você já possui as peças. Use os NOMES e PREÇOS reais da sua zona e NÃO invente produtos.

{tom de abordagem}. VOCÊ É: {biografia do vendedor}.

A.7 Conversa (lado do vendedor)

Contexto de cada turno: o catálogo real da zona ({produtos com nome e preço}) e a fala do cliente.
Resposta: `{"fala": "..."}`
Os tons de abordagem são: **neutro** (responde sem empurrar), **incisivo** (pressiona e insiste), **calmo** (sugere sem pressionar) e **caloroso** (elogia e cria conexão).