1.5 JUSTIFICATIVA

A relevância desta pesquisa parte de um problema já validado pela primeira geração de assistentes virtuais em IES: a sobrecarga dos canais de atendimento e a natural necessidade da comunidade acadêmica por respostas rápidas. Uma universidade de médio a grande porte pode processar milhares de interações mensais, muitas das quais são consultas repetitivas que desviam horas de trabalho de servidores qualificados de suas funções estratégicas. Estudos em instituições de porte similar, como o da *Nguyen et al. (2021)* com o NEU-Chatbot, indicam que sistemas de automação podem absorver até 80% desse volume. Em termos de produtividade, a automação de tarefas libera o capital humano para se concentrar em casos complexos e no planejamento estratégico, alinhando-se à busca por uma gestão universitária mais eficiente.

Contudo, **não basta adotar automação: é preciso garantir a qualidade do serviço entregue**. As métricas tradicionais, desenvolvidas para serviços presenciais ou genéricos, não capturam com precisão as interações específicas dos chatbots em ambientes acadêmicos. Dimensões como aceitação pelo usuário, aderência a normativos institucionais e facilidade de atualização são pouco contempladas nos modelos clássicos. Assim, a relevância desta pesquisa não está em provar a necessidade da automação, mas em **propor derivações a métricas consolidadas de avaliação, de modo a adequá-las ao atendimento universitário**.

É nesse ponto que a pesquisa se justifica academicamente: ao adaptar e testar métricas de qualidade em um caso de alto fluxo de atendimentos, ela contribui para preencher uma lacuna na literatura, oferecendo evidências empíricas sobre quais dimensões são mais críticas no contexto das IES. A viabilidade da proposta reside na premissa de que, **ao avaliar o protótipo multiagente sob a ótica da qualidade do serviço**, gestores terão subsídios mais sólidos para decidir pela implantação ou não dessas soluções, indo além da viabilidade técnica ou econômica.

Finalmente, no que tange à competitividade institucional, um atendimento 24/7, preciso e bem avaliado melhora a experiência do estudante, contribui para as estratégias de permanência e fortalece a imagem da IES em um cenário cada vez mais competitivo. O impacto esperado é um círculo virtuoso: **a melhoria na qualidade do serviço prestado via chatbot reforça a eficiência operacional, libera tempo estratégico e aumenta a confiança na gestão universitária**.

1.6 DELIMITAÇÃO DO ESCOPO

Para delinear com precisão o campo de atuação desta pesquisa, torna-se imperativo distinguir dois conceitos centrais: a arquitetura multiagente e o sistema multiagente. A primeira representa o modelo conceitual, a estrutura abstrata que define os papéis dos agentes e suas regras de interação, constituindo a contribuição de design desta pesquisa, cuja metodologia de construção é detalhada entre as seções 3.1 e 3.3.

O sistema multiagente, por sua vez, é a materialização dessa arquitetura: o protótipo executável que serve como prova de conceito e cujo desempenho constitui o objeto central da validação empírica, a ser discutida nos resultados desta pesquisa.

1.6.1 Delimitação Contextual

O escopo contextual desta pesquisa é deliberadamente focado no ecossistema de ensino de Santa Catarina, tendo a Universidade Federal de Santa Catarina (UFSC) como principal ambiente para coleta de dados, desenvolvimento e validação do SMA. Essa escolha se sustenta em três pilares estratégicos:

- Viabilidade de acesso: A proximidade com o objeto de estudo possibilita uma análise aprofundada dos processos institucionais e o acompanhamento direto dos ciclos de prototipagem e avaliação.
- Complexidade representativa: Como universidade federal de grande porte, a UFSC reúne demandas de atendimento diversificadas, oferecendo um cenário representativo dos desafios enfrentados por outras IES.
- Aderência a fomentos regionais: A escolha dialoga com iniciativas de desenvolvimento científico e tecnológico em Santa Catarina, ampliando a relevância regional da pesquisa.

1.6.2 Delimitação Temática e Tecnológica

No recorte temático, a investigação concentra-se em demandas de atendimento recorrentes, entendendo “atendimento”, para fins deste estudo, como duas categorias de tarefas: recuperação de informações objetivas e automação de processos. Em ambos os casos, o critério delimitador é a verificabilidade, ou seja, a possibilidade de fundamentar cada resposta ou ação em documentos e fluxos institucionais. Permanecem fora do escopo as interações de natureza subjetiva, emocional ou que demandem julgamento humano complexo.

Tecnologicamente, exclui-se o treinamento de modelos de linguagem a partir do zero, mantendo-se a possibilidade de realizar *soft fine-tuning* em modelos pré-existentes. A contribuição se concentra na orquestração e avaliação de agentes especializados, desenvolvidos a partir de ferramentas e frameworks predominantemente open source.

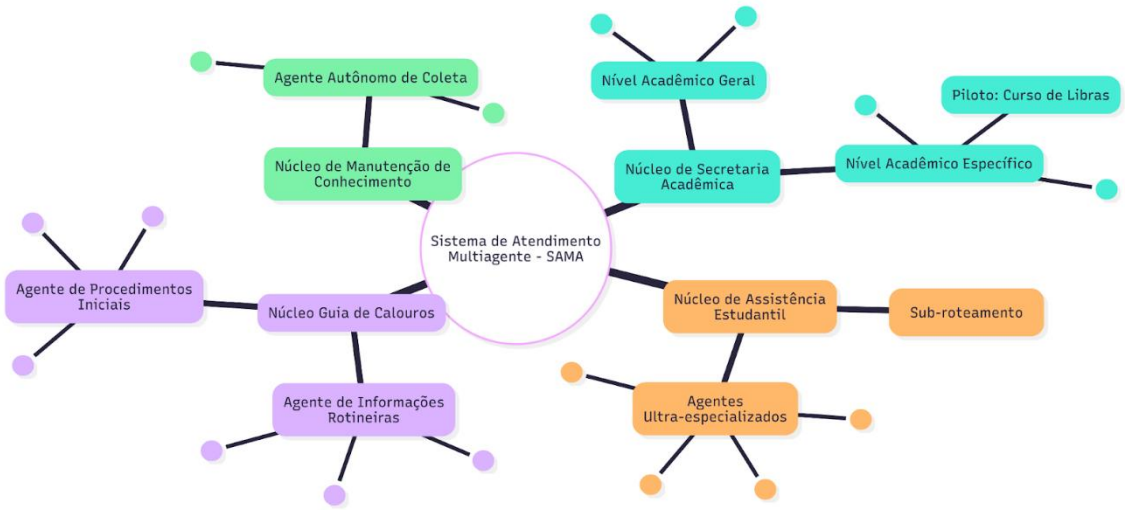
Embora APIs proprietárias possam ser empregadas pontualmente, a estrutura-base, passível de clonagem e adaptação, será aberta e documentada.

Conforme discutido, a entrega tecnológica desta pesquisa não se configura como um manual técnico, mas como um modelo conceitual (a arquitetura) cujo desempenho foi testado por meio de um protótipo (o sistema). A contribuição central do estudo reside em gerar evidências sobre a precisão, os desafios operacionais e a escalabilidade da abordagem multiagente, oferecendo subsídios para que a gestão universitária avalie sua aplicabilidade.

1.6.3 Conjunto Inicial de Agentes Inteligentes

Para os fins desta prova de conceito, determina-se um conjunto inicial de núcleos de agentes, entendidos como agrupamentos de entidades inteligentes especializadas em domínios críticos e representativos do atendimento universitário. A definição desses núcleos, embora informada pela análise de demandas recorrentes (Seção 3.1), permanece flexível, permitindo a decomposição ou agregação de agentes conforme os ciclos iterativos de desenvolvimento indicarem necessário.

Figura 1: Arquitetura Conceitual dos Núcleos de Agentes do SAMA



Fonte: Elaboração Própria (2025).

Este ecossistema de agentes foi desenhado para testar a hipótese de que a decomposição da complexidade em múltiplos níveis de especialização geral, específico de curso e de processo resulta em um atendimento mais preciso e robusto. A avaliação da colaboração entre e dentro desses núcleos constituirá o cerne da análise empírica desta pesquisa.

1.6.5 Delimitação Computacional

Para assegurar que as soluções propostas sejam viáveis mesmo em contextos com infraestrutura limitada, o desenvolvimento e os testes foram realizados em ambiente local, utilizando recursos computacionais pessoais do autor (Tabela 1).

Exclui-se do escopo deste estudo o uso de infraestrutura de pesquisa com múltiplos servidores ou de serviços em nuvem de alto custo.

Tabela 1 – Especificações do Ambiente de Desenvolvimento

Componente	Especificação
Processador (CPU)	AMD Ryzen 7 5600x
Memória RAM	48 GB
Placa de Vídeo (GPU)	NVIDIA GeForce RTX 3060 - VRAM 12 GB
Conexão de Internet	Upload médio de 300 Mbps

Fonte: Elaboração própria (2025).

Essas especificações não configuram requisito mínimo para replicação da metodologia, representando apenas o ambiente empírico utilizado nesta pesquisa. O escopo não abrange o dimensionamento de recursos computacionais nem análise comparativa de desempenho entre configurações alternativas.

Embora a capacidade de atendimento simultâneo dependa de variáveis como a complexidade dos fluxos, o uso de modelos locais ou em nuvem e as configurações do *n8n*, estima-se que o ambiente experimental utilizado possa sustentar centenas de sessões concorrentes quando integrado a *LLMs* via API, e entre três e cinco sessões simultâneas no caso de execução local de modelos de 7B parâmetros quantizados em Q4. Essa estimativa é indicativa e poderá variar conforme a otimização dos fluxos.

Adicionalmente, está no escopo desta pesquisa a utilização de diferentes modelos de LLM no mesmo Sistema Multiagente. Essa abordagem permite alinhar o custo computacional e a complexidade da tarefa ao modelo mais adequado: um agente orquestrador ou responsável por processar informações de alta complexidade, como a análise da estrutura curricular de um curso, pode demandar um modelo de maior desempenho; enquanto um agente que responde a consultas simples, como o cardápio do restaurante universitário, pode manter desempenho satisfatório com um modelo mais modesto.