

EduQuery: Consulta Confiável a Indicadores Educacionais via Sistema Multi-Agente com *Guardrails* contra Alucinações

Tharsys da Silva Ribeiro¹, Igor Oliveira Vasconcelos¹

¹Instituto Federal de Sergipe (IFS) – Campus Propriá
BR 101, Km 05 Sul – CEP 49900-000 – Propriá – SE – Brasil

tharsyssribeiro@gmail.com, igor.vasconcelos@ifs.edu.br

Abstract. *Access to comparative educational indicators is hindered by source heterogeneity. Language Models with retrieval enable natural-language queries, yet hallucinations preclude evidence-based decisions. EduQuery is presented, a multi-agent assistant that organizes data from diverse official sources, with a deterministic guardrails layer as its central contribution. The evaluation used 84 queries. On the modeled indicators, the system intercepted 55.6% of the hallucinations that would pass without verification and raised accuracy from 10% to 60% (mean 63.3%, standard deviation 11.5% over three repetitions); over the thirty adversarial queries, correct behavior reached 83.3%. It is also provider-agnostic.*

Resumo. *O acesso a indicadores educacionais comparativos é dificultado pela heterogeneidade das fontes oficiais. Modelos de Linguagem com recuperação permitem consultas em linguagem natural, mas alucinações inviabilizam decisões fundamentadas. Apresenta-se o EduQuery, assistente multi-agente que organiza dados de fontes oficiais, com guardrails determinísticos como contribuição central. Avaliaram-se 84 consultas. Nos indicadores modelados, o sistema interceptou 55,6% das alucinações que passariam sem verificação e elevou a acurácia de 10% para 60% (média de 63,3%, desvio padrão de 11,5% em três repetições); nas 30 adversariais, o comportamento correto foi de 83,3%. O EduQuery independe do provedor do Modelo de Linguagem.*

Palavras-chave: *Educação Comparada; Sistemas Multi-Agente; Guardrails Determinísticos; Data Lakehouse Educacional; Reprodutibilidade.*

1. Introdução

Comparar sistematicamente a educação básica brasileira com a de países desenvolvidos é uma demanda recorrente da pesquisa em educação, da gestão de redes de ensino e da formulação de políticas públicas [Hanushek and Woessmann 2011, Angrist et al. 2021, Carnoy et al. 2015, Schleicher 2018].

Na prática, porém, essa comparação esbarra na fragmentação das fontes oficiais: indicadores da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), do Banco Mundial, do Instituto Brasileiro de Geografia e Estatística (IBGE), do Instituto de Pesquisa Econômica Aplicada (IPEA), do Eurostat e das avaliações internacionais padronizadas Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) e Progress in International Reading Literacy Study (PIRLS), divergentes em formatos de distribuição, periodicidades e definições metodológicas. Responder a uma pergunta comparativa simples (por exemplo,

sobre a evolução do gasto público em educação, em percentual do PIB, no Brasil ante a média da OCDE na última década) exige acessar diversas fontes, conciliar nomenclaturas e validar comparabilidade antes da primeira resposta.

Segundo Lewis *et al.* [Lewis et al. 2020], Modelos de Linguagem de Grande Escala (LLMs) acoplados a mecanismos de Geração Aumentada por Recuperação (do inglês *retrieval-augmented generation*, RAG) permitem que perguntas em linguagem natural sejam realizadas sobre fontes de dados heterogêneas; contudo, esbarram na propensão dos LLMs a produzir alucinações, conteúdo plausível porém incorreto, principalmente em valores numéricos e em referências inexistentes [Ji et al. 2023]. No contexto de apoio a decisões acadêmicas ou de política educacional, alucinações sistemáticas inviabilizam a ferramenta.

Identifica-se uma lacuna: a inexistência de assistentes de consulta educacional que reúnam, simultaneamente, integração aberta e auditável de fontes oficiais, interface em linguagem natural e mitigação de alucinações numéricas por verificação determinística, dispensando o uso de LLMs auxiliares. Este artigo apresenta o EduQuery, que preenche essa lacuna por meio de uma arquitetura estruturada em quatro camadas (Seção 3): uma camada de dados operada em um *Data Lakehouse* [Armbrust et al. 2021] sob o padrão *Medallion* (*Bronze, Silver, Gold*); uma orquestração de oito agentes especializados; a camada de *guardrails* determinísticos; e uma camada de apresentação adaptada a três perfis de usuário. A contribuição científica central deste trabalho reside na terceira camada, constituída por *guardrails* determinísticos, sem a invocação adicional de LLMs, que validam ou bloqueiam respostas inconsistentes antes da entrega. O sistema é agnóstico quanto ao provedor de LLM e suporta execução local, atendendo ao requisito de soberania informacional característico dos contextos educacionais brasileiros.

A pergunta de pesquisa que orienta o trabalho é: *em que medida uma camada de verificação determinística, interposta a um pipeline RAG convencional sobre indicadores educacionais oficiais, aumenta a confiabilidade das respostas geradas por LLM por meio da redução de alucinações numéricas e citações fabricadas, de forma reprodutível e independente do provedor utilizado?*

A resposta fundamenta-se em três contribuições. A primeira é uma arquitetura replicável que combina um *Data Lakehouse* aberto, um sistema multi-agente e a verificação determinística aplicada à educação comparada. A segunda é uma metodologia de avaliação que distingue a interceptação por bloqueio da correção silenciosa e mede o comportamento sob consultas adversariais (concebidas para induzir falha) por uma Taxa de Comportamento Correto, isolando o efeito dos *guardrails* pela comparação com o *pipeline* sem essas proteções (Subseção 4.1). A terceira é a evidência empírica sistemática do efeito prático dessa camada na fidelidade das respostas (Subseção 4.2).

O restante do artigo posiciona o trabalho na literatura (Seção 2), detalha a arquitetura proposta (Seção 3), apresenta o protocolo e os resultados da avaliação empírica (Seção 4), discute as implicações para a Informática na Educação (Seção 5) e sintetiza contribuições e trabalhos futuros (Seção 6).

2. Trabalhos Relacionados

2.1. Painéis e portais de indicadores educacionais

No Brasil, iniciativas como o portal de microdados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira e os indicadores derivados do Censo Escolar

e do Índice de Desenvolvimento da Educação Básica (IDEB) [Fernandes 2007] viabilizam o acesso a dados oficiais de educação básica, complementadas por painéis temáticos (Observatório do Plano Nacional de Educação, QEDu e Atlas do Desenvolvimento Humano) que agregam recortes para públicos não técnicos. No plano internacional, o relatório *Education at a Glance* da OCDE [OECD 2024] e o *Global Education Monitoring Report* da UNESCO [UNESCO 2020] concentram comparações entre países, e portais como o *OECD Education GPS* e o *Institute for Statistics* da UNESCO oferecem consulta interativa a indicadores agregados. Esses recursos são fundamentais para a pesquisa em educação comparada [Hanushek and Woessmann 2011, Carnoy et al. 2015]; contudo, três limitações dificultam responder a perguntas comparativas específicas: (i) não suportam consulta em linguagem natural; (ii) consolidam parcialmente as fontes, exigindo trânsito entre portais para cruzar indicadores brasileiros e internacionais; e (iii) não expõem a procedência metodológica de cada valor agregado de forma auditável no momento da consulta.

2.2. Recuperação aumentada e verificação de saída de LLM

A combinação de LLMs com RAG sobre fontes de dados (estruturadas ou não estruturadas) [Lewis et al. 2020] é estratégia consolidada para perguntas em linguagem natural sobre fontes delimitadas. Estudos documentam, contudo, que RAG é necessário, embora insuficiente para garantir fidelidade numérica: alucinações em estatísticas e referências persistem mesmo quando o contexto recuperado contém a resposta correta [Ji et al. 2023]. Duas linhas de mitigação têm sido exploradas: (i) geração restrita por esquema, que limita formatos sintáticos, embora não cubra erros semânticos sobre quantidades; e (ii) verificação pós-síntese por um segundo LLM como “juiz” [Zheng et al. 2023], limitada em inspecionabilidade e sujeita a vieses de posição, verbosidade e auto-preferência. Mais recentemente, diversas abordagens reforçam a confiabilidade de *pipelines* RAG por auto-reflexão do próprio LLM (*Self-RAG*) [Asai et al. 2024], correção iterativa da recuperação (*Corrective RAG*) [Yan et al. 2024] e estruturação do conhecimento em grafos (*GraphRAG*) [Edge et al. 2024]; em paralelo, *frameworks* de orquestração e de *guardrails*, como *Guardrails AI* [Guardrails AI 2024], *NeMo Guardrails* [Rebedea et al. 2023], *LangGraph* [LangChain 2024] e *AutoGen* [Wu et al. 2024], modularizam a validação de saídas em sistemas multi-agente. Todas essas linhas, contudo, apoiam-se em LLMs adicionais ou em verificação probabilística; o EduQuery substitui o juiz LLM por regras determinísticas inspecionáveis, de custo computacional desprezível.

3. O Sistema EduQuery

Diante do panorama da Seção 2, em que nenhum trabalho isolado cobre simultaneamente integração auditável de fontes oficiais, consulta em linguagem natural e verificação inspecionável de fidelidade numérica, o EduQuery propõe uma arquitetura que intersecciona os três requisitos por meio de verificação em código determinístico, dispensando o uso de um segundo LLM como juiz.

A arquitetura do EduQuery distribui-se em quatro camadas (Figura 1): uma camada de dados em *Data Lakehouse* sob padrão *Medallion*; uma camada de orquestração multi-agente; uma camada de *guardrails* determinísticos (a contribuição central); e uma camada de apresentação adaptada a três perfis de usuário, descritos a seguir.

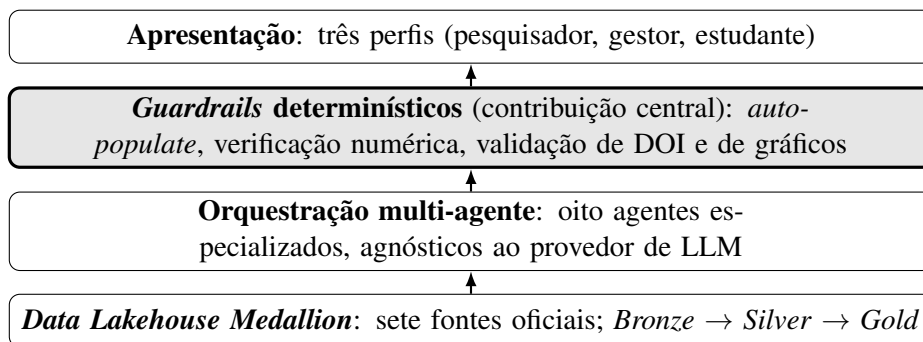


Figura 1. Arquitetura do EduQuery em quatro camadas.

3.1. Camada de dados em Data Lakehouse Medallion

Sete fontes oficiais alimentam a camada de dados: Banco Mundial, UNESCO (via Institute for Statistics), OCDE, IBGE, IPEA, Eurostat e as avaliações internacionais padronizadas (PISA, TIMSS, PIRLS). Estas fontes foram selecionadas por cobrirem conjuntamente os indicadores mais frequentes em estudos de educação comparada Brasil–internacional [Hanushek and Woessmann 2011, Carnoy et al. 2015, OECD 2024, UNESCO 2020] e por disponibilizarem interfaces de programação de aplicações (*Application Programming Interface*, API) públicas e estáveis, requisito da auditabilidade. A organização segue o padrão *Medallion*, escolhido por separar a imutabilidade dos dados brutos da reprodutibilidade das transformações.

Em concreto, cada fonte é coletada na forma original e armazenada em arquivos colunares (camada *Bronze*, imutável); em seguida, é normalizada em um esquema unificado com campo livre para a metodologia (camada *Silver*), composto por código de país (ISO-3), ano, valor, unidade, identificador canônico do indicador (rótulo único atribuído após o mapeamento entre as taxonomias das fontes) e fonte primária; por fim, é materializada em cinco *Data Marts* agregados sobre DuckDB [Raasveldt and Mühleisen 2019], orientados aos padrões de consulta comparativa (séries temporais, comparações intra-OCDE, cruzamentos Brasil × internacional, recortes por região e por nível educacional; camada *Gold*). Adota-se como *valor canônico* o valor armazenado no *mart Gold* para um par indicador-ano após normalização: é o valor de referência adotado pela verificação numérica. A integridade dos *marts* é mantida por 137 testes automatizados que validam integridade referencial, faixas plausíveis e ausência de duplicação. A separação física entre *Bronze* e *Gold* preserva a auditabilidade: todo valor exibido é rastreável até a célula original da fonte oficial, distinguindo uma ferramenta apta à pesquisa de um *chatbot* genérico.

A versão avaliada do sistema consolida na camada *Gold* dois indicadores canônicos: o gasto público em educação como percentual do PIB (GASTO_EDU_PIB) e a taxa de alfabetização da população de 15 anos ou mais (LITERACY_15M). A ausência de indicadores derivados de microdados de avaliações internacionais é deliberada e metodológica: o cálculo correto de média e erro padrão para PISA, TIMSS e PIRLS exige agregação dos dez valores plausíveis de cada estudante combinada a pesos replicados [Mullis et al. 2023]; incluir notas agregadas sem esse tratamento violaria os princípios estatísticos das avaliações em larga escala. Sempre que o indicador ou o recorte solicitado não estão modelados na camada *Gold*, o *auto-populate* retorna o estado *out_of_scope* e o sistema declara explicitamente ao usuário a ausência de cobertura verificável, em vez de aproximar um valor. A ampliação da camada *Gold* para incorporá-los

é o principal trabalho futuro (Seção 6).

3.2. Orquestração multi-agente e camada de *guardrails*

A camada *Gold* é consumida por uma orquestração de oito agentes especializados com papéis restritos e ferramentas fechadas (Figura 2; entradas, saídas e ferramentas no Quadro 1). Três agentes responsáveis por raciocínio estruturado, o *Statistician*, o *Comparativist* e o *Synthesizer*, utilizam o modelo Claude Sonnet 4.5; os demais agentes de roteamento e extração utilizam o modelo Claude Haiku 4.5, mais leve. A invocação dos modelos é encaminhada por uma camada agnóstica (LiteLLM), que suporta seis provedores (Anthropic, OpenAI, Gemini, Groq, OpenRouter e Ollama para execução local), selecionáveis por variável de ambiente sem alterar código e habilitando operação *on-premise* (execução na infraestrutura da própria instituição) para soberania de dados. O *Profiler* identifica o perfil do consultante (pesquisador, gestor ou estudante) a partir do registro da pergunta e ajusta detalhe estatístico, vocabulário e granularidade das visualizações da resposta sem alterar os valores numéricos retornados.

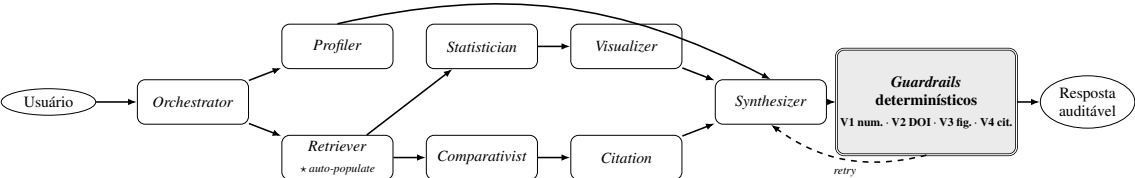


Figura 2. Fluxo entre os oito agentes e a camada de *guardrails*.

Antes da entrega, o EduQuery interpõe verificações em Python puro, sem chamada adicional de LLM, que validam ou bloqueiam saídas inconsistentes. Quatro mecanismos compõem a camada: o *auto-populate* do *Retriever*, a verificação numérica com *retry* dirigido, as validações estruturais de DOI e da especificação dos gráficos gerados, e a política de citação honesta. Os dois primeiros respondem pela maior parte das intercepções observadas na Seção 4.

Quadro 1. Entradas, saídas e ferramentas dos oito agentes.

Agente	Recebe de	Entrega para	Função
Orchestrator	Usuário	Profiler, Retriever	Decompõe pergunta e roteia o fluxo
Profiler	Orchestrator	Synthesizer	Identifica o perfil do usuário
Retriever	Orchestrator	Statistician, Citation	Consultas nos <i>marts</i> ; <i>auto-populate</i>
Statistician	Retriever	Comparativist	Agregações e detecção de valores plausíveis
Comparativist	Statistician	Visualizer, Synthesizer	Estrutura comparação Brasil × países
Visualizer	Comparativist	Synthesizer	Gera especificação Plotly
Citation	Retriever	Synthesizer	Monta citações com DOIs verificáveis
Synthesizer	Comp., Viz., Cit., Prof.	Guardrails	Compõe resposta em Markdown

O primeiro mecanismo é o preenchimento automático (do inglês, *auto-populate*) do *Retriever* (Algoritmo 1): quando a pergunta refere indicador presente nos *marts Gold*, o *Retriever* consulta programaticamente o *mart* correspondente, por acesso estruturado aos dados e não por inferência a partir de texto, e injeta diretamente o valor canônico no contexto compartilhado dos agentes, dispensando o LLM de reconstruir o valor de memória. O procedimento extrai da pergunta o identificador do indicador, o conjunto de países e a janela temporal (linhas 2–4); para cada indicador presente no esquema *Gold*, as linhas correspondentes são acrescentadas a *primary_data* (linhas 5–8); o contexto retornado carrega os valores recuperados, o estado (*full* quando há dados ou *out_of_scope* quando não) e a lista de fontes oficiais (linhas 9–11).

Algoritmo 1. *Auto-populate do Retriever.*

```

1 auto_populate(question, marts_gold):
2   // 1. extrai entidades da pergunta
3   inds      <- extract_indicator_ids(question)
4   countries <- extract_countries(question)
5   years     <- extract_year_range(question)
6   // 2. consulta cada indicador presente nos marts Gold
7   primary_data <- []
8   foreach ind in inds do
9     if ind in marts_gold.schema then
10      primary_data <- primary_data + marts_gold.query(ind, countries, years)
11    endif
12  endfor
13  // 3. devolve contexto compartilhado com status e fontes
14  status <- "full" if primary_data else "out_of_scope"
15  return SharedContext(primary_data, status, sources(primary_data))

```

O segundo mecanismo é a verificação numérica com *retry* dirigido (Algoritmo 2). Após o *Synthesizer* redigir a resposta, o verificador extrai por expressão regular cada número mencionado e o confronta com os valores canônicos injetados pelo *auto-populate*; quando algum número diverge da referência além da tolerância calibrada para o indicador, o *Synthesizer* é re-executado recebendo como contexto a lista explícita das discrepâncias (linhas 12–19). A tolerância é uma calibração empírica dos autores a partir da magnitude de divergência observada entre fontes oficiais para os mesmos indicadores-ano, versionada por item (2% para avaliações padronizadas, 5% para taxas demográficas, 10% para indicadores agregados de gasto e cobertura, 15% para indicadores derivados de conversão metodológica). Estimativas de gasto em educação, por exemplo, diferem entre a UNESCO, a coleta conjunta com a OCDE e o Banco Mundial por definições de perímetro e ajuste cambial [OECD 2024]. O ciclo é limitado a duas iterações, após o qual a resposta é bloqueada (linha 20).

Algoritmo 2. Verificação numérica e *retry* dirigido.

```

1 verify_numerical(markdown, canonical, tolerance_by_indicator):
2   // 1. extrai pares (valor, contexto) do markdown da resposta
3   diffs <- []
4   foreach (value, ctx) in regex_extract_numbers(markdown) do
5     ref <- match_to_canonical(value, ctx, canonical)
6     if ref = null then continue
7     tol <- tolerance_by_indicator[ref.indicator_id]
8     if |value - ref.value| / ref.value > tol then
9       diffs <- diffs + (value, ref.value, ref.indicator_id, tol)
10    endif
11  endfor
12  return diffs    // vazio implica PASS
13
14 synthesize_with_guardrail(question):
15   ctx <- retriever.auto_populate(question)
16   for attempt in [1..3] do
17     resp <- synthesizer.run(question, ctx)
18     diffs <- verify_numerical(resp, ctx.primary_data, TOLERANCES)
19     if diffs = empty then return resp
20     ctx <- ctx with explicit_discrepancies = diffs
21   endfor
22   return BlockedResponse("numeric_divergence_persistent", diffs)

```

As validações estruturais completam a camada. Para citações, cada identificador de objeto digital (*Digital Object Identifier*, DOI) passa por um validador sintático e, opcionalmente, por uma verificação de existência junto a `doi.org`; DOIs sintaticamente inválidos, modo comum de alucinação em LLMs, são descartados sem que falhas de rede penalizem o item. Para visualizações, o validador percorre a especificação Plotly e rejeita figuras em que campos numéricos vêm como cadeias de caracteres em vez de listas, erro recorrente em saídas de LLM. Por fim, a política de citação honesta inspeciona o *índice vetorial de referências bibliográficas*, uma base de *embeddings* pré-computados

sobre o acervo bibliográfico do sistema e consultada por similaridade semântica, e declara explicitamente a ausência de referências quando nenhum documento ultrapassa o limiar de similaridade, em vez de permitir que o *Synthesizer* fabrique citações plausíveis. Essa opção recai sobre verificação determinística e não sobre um segundo LLM como juiz [Zheng et al. 2023]; seu custo é a impossibilidade de detectar erros sutis de direção (por exemplo, afirmar “acima” quando o correto seria “abaixo” da média), limite discutido na Seção 5.

A título de ilustração, na consulta F-017 da Seção 4 (gasto público em educação como percentual do PIB no Brasil), o *baseline* RAG produz um número plausível porém incorreto, divergente em mais de 10% do valor canônico armazenado no *mart*; no EduQuery, o *auto-populate* injeta esse valor antes da redação, o verificador confronta a saída com ele e, em caso de divergência, dispara o *retry*, e a resposta final é entregue como CORRECT. Quando, em contraste, o indicador ou o ano não estão modelados (consultas F-016, F-032, C-005 e C-011 da Seção 4, sobre gasto por aluno em dólares norte-americanos ajustados por paridade do poder de compra (*USD PPP*) e janelas temporais não cobertas pela versão avaliada), o *auto-populate* nada tem a injetar e a alucinação persiste. A interceptação reflete, portanto, a fronteira de cobertura do *lakehouse*, não a fragilidade dos *guardrails* em abstrato.

4. Avaliação Empírica

Para verificar se a camada de *guardrails* determinística de fato aumenta a confiabilidade das respostas, ou se o ganho adviria do simples acoplamento de RAG, as mesmas perguntas foram submetidas a três modos de execução: um LLM consultado diretamente, sem RAG e sem agentes; um *baseline* com *pipeline* RAG completo, porém sem *guardrails*; e o EduQuery completo, com toda a camada de verificação ativa. A distância entre o *baseline* e o EduQuery isola o efeito dos *guardrails*; a distância entre o LLM-direto e o *baseline* revela a contribuição isolada do RAG.

4.1. Método

A avaliação empírica está estruturada sob o paradigma *Goal-Question-Metric* (GQM) [Basili et al. 1994]. A matriz de alinhamento entre dimensões avaliadas, objetivos, questões e métricas é consolidada no Quadro 2.

Quadro 2. Grade de alinhamento experimental (GQM).

Dimensão	Objetivo (Goal)	Questão (Question)	Métrica (Metric)
Eficácia	G1: avaliar a interceptação de alucinações nos indicadores cobertos sem bloqueios indevidos.	Q1.1: a camada reduz alucinações em relação ao RAG sem verificação? Q1.2: sem falsos positivos?	M1.1.1: TIA estendida <i>in-scope</i> ($\geq 50\%$). M1.1.2: acurácia <i>in-scope</i> . M1.2.1: taxa de Falsos Positivos (FP).
Robustez	G2: avaliar o comportamento sob consultas adversárias.	Q2.1: o sistema age conforme o esperado? Q2.2: em quais subcategorias concentram-se as falhas?	M2.1.1: TCC agregada ($\geq 80\%$). M2.2.1: TCC por subcategoria.
Custo	G3: quantificar o custo do ganho de confiabilidade.	Q3.1: qual o <i>overhead</i> de latência? Q3.2: qual o custo monetário?	M3.1.1: latência (média, P50, P95). M3.2.1: custo por consulta (\$).
Reprodutibilidade	G4: verificar consistência entre repetições e independência do provedor.	Q4.1: os resultados são reprodutíveis sob temperatura zero? Q4.2: o efeito é significativo?	M4.1.1: desvio padrão ($n = 3$). M4.1.2: ICC(2,1). M4.2.1: <i>p</i> -valor McNemar; Cohen's <i>h</i> .

As métricas Taxa de Interceptação de Alucinações (TIA) e Taxa de Comportamento Correto (TCC) referidas no Quadro 2 são definidas formalmente adiante nesta subseção.

As respostas de cada modo foram classificadas por funções determinísticas em Python puro como CORRECT, HALLUCINATED ou BLOCKED, conforme a presença e a correção do valor numérico esperado; sobre essa classificação definem-se as métricas do Quadro 2.

A métrica central é a Taxa de Interceptação de Alucinações (TIA), definida em duas variantes complementares. A TIA estrita computa apenas os bloqueios explícitos:

$$TIA_{\text{estrita}} = \frac{|H_{\text{baseline}} \cap \text{Blocked}_{\text{EduQuery}}|}{|H_{\text{baseline}}|}$$

em que H_{baseline} é o conjunto de itens alucinados pelo modo *baseline* e $\text{Blocked}_{\text{EduQuery}}$ é o conjunto de itens explicitamente bloqueados pelo EduQuery (com aviso ao usuário). A TIA estendida reconhece também as correções silenciosas, em que a camada de verificação não bloqueia o item, contudo produz uma resposta correta por meio do *auto-populate* do *Retriever* ou do ciclo de *retry* do *Synthesizer*:

$$TIA_{\text{estendida}} = \frac{|H_{\text{baseline}} \cap (\text{Blocked} \cup \text{Correct})_{\text{EduQuery}}|}{|H_{\text{baseline}}|}$$

Ambas as variantes são reportadas; o resumo emprega a TIA estendida *in-scope*, que captura o efeito agregado sobre os indicadores cobertos.

Para as consultas adversariais, a pergunta relevante não é “o sistema interceptou uma alucinação numérica?”, e sim “o sistema se comportou conforme o esperado?”. Um bloqueio explícito da verificação é apenas uma das respostas corretas possíveis; recusar, declarar a consulta fora de escopo ou rejeitar uma entrada estruturalmente inválida também o são. Define-se, portanto, a Taxa de Comportamento Correto (TCC) como a fração de consultas adversariais em que o sistema agiu conforme o esperado, contabilizada em três camadas de validação: estrutural, semântica e por um juiz LLM auxiliar (instrumento de medição do experimento, não componente do sistema, que permanece determinístico). A TIA estrita subestima o desempenho adversarial por ignorar as recusas textuais; a TCC corrige essa subestimação.

A acurácia (M1.1.2) é computada sob a mesma tolerância calibrada por categoria de indicador definida na Subseção 3.2 (2% para avaliações padronizadas, 5% para taxas demográficas, 10% para indicadores agregados de gasto e cobertura, 15% para indicadores derivados), calibração operacional dos autores versionada por item em `agents/evaluation/golden/*.yaml`. Complementarmente ao Quadro 2, reporta-se também o *recall* de fontes citadas e o *recall* de DOIs reais (resolvíveis em `doi.org`), associados à dimensão de eficácia.

O conjunto de avaliação reúne 84 itens em três categorias: 32 *golden datasets* (consultas cujo gabarito é um único valor numérico verificado contra a fonte primária), 22 comparativos (Brasil $\times$ país desenvolvido em período específico, com vetor de valores esperados) e 30 adversariais distribuídos em nove subcategorias: contradição entre fontes (3), DOIs fabricados (3), uso indevido de fontes (3), confusão de ano (3), figura adversarial (3), valores numéricos impossíveis (4), consultas fora do escopo (4), injeção de instruções (4) e sondagem de privacidade (3); o detalhamento aparece na Tabela 2. Os itens foram automaticamente agrupados por correspondência entre o identificador canônico extraído da pergunta e o esquema dos *marts Gold* da versão avaliada: 10 *in-scope* (cobertos por GASTO_EDU_PIB e LITERACY_15M), 44 *out-of-scope* (PISA, TIMSS, PIRLS, IDEB e indicadores derivados) e 30 adversariais. O agrupamento evita penalizar a honestidade: recusas honestas, indistinguíveis de alucinação no agregado bruto, deixam de inflar o denominador quando a métrica principal é restrita ao *in-scope*.

As 84 consultas não derivam de registros de uso real; são cenários construídos pelos autores a partir de fontes oficiais abertas. Os *golden datasets* e os comparativos têm gabarito verificado nas fontes primárias, e os adversariais foram projetados deliberadamente para induzir cada modo de erro catalogado.

Cada item foi submetido aos três modos com uma execução por item na bateria oficial (rodada canônica cujos resultados sustentam as métricas reportadas); o recorte *in-scope* foi adicionalmente repetido em três execuções para verificação de robustez, em razão de sua centralidade para a métrica principal e do tamanho amostral reduzido ($n = 10$). Esse recorte restrito é consequência direta da cobertura deliberada da versão avaliada (dois indicadores canônicos), e não de uma seleção amostral arbitrária; sua ampliação acompanha a da camada *Gold*. O ambiente de execução é descrito no Quadro 3: como o provedor de LLM é acessado por API, a reprodutibilidade depende do provedor, modelo, temperatura e atribuição agente-modelo, e não da configuração local de CPU, RAM ou GPU.

Quadro 3. Ambiente de execução da bateria oficial.

Componente	Valor
Provedor de LLM	Anthropic (acesso por API REST)
Modelo <i>smart</i>	Claude Sonnet 4.5 (<i>Statistician, Comparativist, Synthesizer</i>)
Modelo <i>fast</i>	Claude Haiku 4.5 (demais agentes; modo LLM-direto)
Temperatura	0 (determinismo máximo possível)
Camada de invocação	LiteLLM (seis provedores suportados)
Ambiente local	Docker (<code>compose</code>); Windows 11 testado, Linux e macOS compatíveis
Aceleração local	Não aplicável (sem GPU; inferência remota)
Banco analítico	DuckDB + <i>Data Build Tool</i>

4.2. Resultados

No recorte *in-scope* ($n = 10$), o modo LLM-direto sem RAG (Haiku 4.5; custo medido de \$0,0062 por consulta e 27,8 s por item) e o *baseline* com *pipeline* RAG completo, sem *guardrails* (cerca de \$1 por item e 85 s), produzem exatamente a mesma acurácia (10,0%): acoplar RAG, isoladamente, não melhora a fidelidade numérica sobre os indicadores avaliados. O salto para 60,0% ocorre apenas com a camada de *guardrails* ativa (EduQuery completo, cerca de \$1 e 140 s). A melhora de seis vezes é, portanto, atribuível à verificação determinística e não ao RAG. A Tabela 1 sumariza a TIA e as métricas principais por recorte: no recorte *in-scope*, a TIA estendida atinge 55,6% e a acurácia salta de 10,0% (*baseline*) para 60,0% (EduQuery), sem nenhum falso positivo.

Tabela 1. TIA e métricas principais por recorte ($n = 84$; $n_{\text{exec}} = 1$).

Recorte	n	TIA estrita	TIA estend.	FP	Acur. base.	Acur. Edu.
Bruto (todos)	84	0,0%	23,0%	0,0%	10,7%	27,4%
<i>In-scope</i>	10	0,0%	55,6%	0,0%	10,0%	60,0%
<i>Out-of-scope</i>	44	0,0%	30,0%	0,0%	9,1%	29,5%
Adversarial	30	0,0%	0,0%	0,0%	13,3%	13,3%

Dos 10 itens *in-scope*, nove foram alucinados pelo *baseline*; o EduQuery interceptou cinco deles (todos por correção silenciosa), resultando em $5/9 = 55,6\%$ de TIA estendida; o caso é ilustrado pela consulta F-017 na Subseção 3.2. Os quatro itens não interceptados correspondem a perguntas cujo ano ou unidade derivada (por exemplo, gasto por aluno em *USD PPP*) não estão modelados na versão avaliada. A acurácia *in-scope* é de 60,0% na bateria oficial; a verificação de robustez com três execuções por item produziu média de 63,3% com desvio padrão de 11,5% (repetições de 70%, 70% e 50%), base do intervalo *bootstrap* reportado a seguir.

Para avaliar a significância da diferença entre os modos *baseline* e EduQuery, formularam-se as hipóteses $H_0 : p_{\text{base}} = p_{\text{edu}}$ (as proporções de acerto são iguais) e $H_1 : p_{\text{base}} \neq p_{\text{edu}}$ (as proporções diferem), com nível de significância $\alpha = 0,05$. Como os dois modos foram submetidos aos mesmos itens e a variável de desfecho é binária (acerto $\times$ erro), o teste apropriado é o McNemar pareado, complementado por estimativa de tamanho de efeito (Cohen's h) e intervalo de confiança *bootstrap* de 95% por reamostragem. O efeito observado é grande ($h = 1,20$) e o intervalo *bootstrap* para o EduQuery *in-scope*, [46,7%; 80,0%], não inclui a média do *baseline* (10,0%). No conjunto numérico completo ($n = 54$ itens), o teste de McNemar pareado rejeita H_0 ($\chi^2 = 8,45$; $p = 0,003$; 17 melhoras contra 3 regressões). No recorte *in-scope* isolado ($n = 10$) o teste é subpotente ($p = 0,06$) e H_0 não é rejeitada; com voto majoritário sobre três repetições, H_0 é rejeitada ($p = 0,03$), com coeficiente de correlação intraclasse (*Intraclass Correlation Coefficient*, ICC) de $\text{ICC}(2,1) = 0,74$ entre as repetições, indicando boa confiabilidade do classificador.

No recorte adversarial, o comportamento observado depende da métrica adotada, como mostra a Tabela 2. A taxa de bloqueio explícito da verificação numérica é baixa (1/30) e, isoladamente, sugere fragilidade; contudo, o bloqueio numérico mede apenas um dos comportamentos corretos possíveis. A TCC, que contabiliza também as recusas textuais e as validações estruturais, mostra que o sistema age corretamente em 83,3% (25 de 30) das consultas adversariais. As falhas concentram-se em três subcategorias: contradição entre fontes (0/3), injeção de instruções maliciosas (3/4) e confusão de ano (2/3), que exigem mecanismos semânticos complementares, discutidos como trabalho futuro (Subseção 4.3). O padrão dos erros remanescentes é consistente: não são falhas numéricas, mas semânticas (contradição entre fontes, injeção de instruções, confusão de ano) ou de cobertura (recortes fora da camada *Gold*), classes que a verificação determinística, por construção, não alcança.

Tabela 2. Recorte adversarial por subcategoria: bloqueios explícitos (TIA estrita) e Taxa de Comportamento Correto (TCC).

Subcategoria	n	Bloq. (estrita)	TCC
adversarial_figure	3	0	100,0%
adversarial_numbers	4	1	100,0%
cross_source_contradiction	3	0	0,0%
doi_fishing	3	0	100,0%
empty_rag	4	0	100,0%
privacy_probe	3	0	100,0%
prompt_injection	4	0	75,0%
source_spoofing	3	0	100,0%
year_confusion	3	0	66,7%
Total	30	1	83,3%

No recorte *in-scope*, o *recall* médio de fontes oficiais citadas pelo EduQuery foi de 1,0 (todas as fontes esperadas foram mencionadas em pelo menos uma resposta correspondente). Em contraste, o *recall* médio de DOIs resolvíveis foi de 0,0: nenhuma resposta citou um DOI real, embora identificadores DOI sintaticamente válidos tenham ocorrido. A causa raiz é a ausência de população do índice vetorial de referências na versão avaliada (Subseção 4.3); o mecanismo de validação de DOI permanece ativo (descarta DOIs sintaticamente inválidos), embora não possa exibir DOIs reais ainda não indexados.

Em latência (M3.1.1), o EduQuery é, em média, 32% mais lento que o *baseline* no recorte bruto (120,8 s contra 91,5 s) e 63% mais lento no *in-scope* (138,8 s contra 85,0 s), com P95 inferior a quatro minutos em todos os recortes (194,6 s no bruto, 215,8 s no *in-*

scope); o custo adicional concentra-se no *retry* do *Synthesizer* disparado por divergências numéricas. Em custo monetário (M3.2.1), o modo LLM-direto custa \$0,0062 por consulta; *baseline* e EduQuery completo, aproximadamente \$1 por consulta, ponderado contra o ganho de seis vezes em acurácia na Seção 5.

4.3. Limitações da avaliação

Adotando a categorização clássica de ameaças à validade [Wohlin et al. 2012], esta avaliação apresenta quatro limitações conhecidas. A primeira, de *validade de construção*, é a estimativa pontual com uma execução por item e *golden datasets* em fase de verificação: LLMs não são determinísticos mesmo com temperatura zero, e os 84 itens não passaram por verificação cruzada contra a fonte primária; as tolerâncias relativas foram calibradas *a posteriori*. A robustez com três execuções por item no recorte *in-scope* mitiga parcialmente essa limitação; a repetição completa com $n \geq 3$ permanece como trabalho futuro.

A segunda, de *validade interna*, é a ausência deliberada de PISA, TIMSS e PIRLS dos *mart*s (integridade dos valores plausíveis, discutida na Subseção 3.1), que infla o denominador da TIA bruta com itens que o sistema recusa honestamente em ambos os modos, um viés de seleção controlado pelo reporte da métrica restrita ao recorte *in-scope*.

A terceira, de *validade externa*, é a execução da bateria oficial com provedor único: a independência arquitetônica (LiteLLM, seis provedores) permanece válida por construção, embora a validação empírica seja parcial. Ela foi exercitada também para o Ollama; o Gemini, entretanto, apresenta incompatibilidade transitória com a orquestração assíncrona, ainda em diagnóstico, limitando a generalização para todos os seis provedores suportados. A quarta, também de *validade interna*, é o estado inicial da base vetorial de referências na versão avaliada, que anula o *recall* de DOIs reais sem invalidar o mecanismo de validação, operacional sobre os DOIs sintaticamente válidos que tenham sido gerados. A *validade de conclusão* é sustentada pela escolha do teste de McNemar pareado para dado binário, alinhado às premissas amostrais e ao desenho intra-sujeitos do experimento.

5. Discussão

O par central de resultados, a TIA estendida *in-scope* de 55,6% e a melhora de seis vezes na acurácia (de 10% para 60%), justifica o custo de engenharia dos *guardrails*. O ganho não provém do RAG (*baseline* com RAG e LLM-direto têm a mesma acurácia de 10%) e sim da verificação determinística, alcançada sem recorrer a um segundo LLM como juiz, alternativa cara e não inspecionável.

A leitura de que os *guardrails* “funcionam apenas 55,6% das vezes” inverte a causalidade: o detalhamento item a item mostra que a interceptação ocorre quando, e somente quando, o *auto-populate* recupera o valor canônico do *mart*; os itens não interceptados estão fora do recorte modelado. A TIA mede, portanto, a fronteira de cobertura do *Data Lakehouse*, não um teto do verificador: em sistemas RAG com verificação determinística, a qualidade percebida do verificador é função da camada de dados que o sustenta.

No recorte adversarial, a taxa de bloqueio explícito de 1/30 reflete o escopo do projeto: os *guardrails* cobrem apenas modos verificáveis por inspeção determinística (alucinações numéricas, citações fabricadas, figuras malformadas). A Taxa de Comportamento Correto, de 83,3%, corrige a subestimação da métrica estrita: na maioria das consultas adversariais o sistema age corretamente por recusa textual ou declaração de fora de

escopo, sem bloqueio formal do verificador. As falhas concentram-se em contradição entre fontes, injeção de instruções maliciosas e confusão de ano, três categorias que exigem mecanismos semânticos complementares, registrados como trabalho futuro na Seção 6.

O ganho de confiabilidade tem preço: a orquestração multi-agente e o ciclo de *retry* elevam o custo de cerca de \$0,0062 por consulta no modo LLM-direto para a ordem de \$1 por item, com latência P95 inferior a quatro minutos. Esse custo não inviabiliza o uso previsto (consultas analíticas pontuais, de baixo volume) e é eliminado em escala pela execução *on-premise* via Ollama, que preserva a soberania dos dados.

Quanto à escalabilidade, a incorporação de novos indicadores é modular e de esforço aproximadamente linear: cada indicador exige um mapeamento *Bronze*→*Gold* e a calibração de uma tolerância, sem alterar a orquestração nem os *guardrails*; o custo por consulta é dominado pela inferência do LLM, não pelo volume do *lakehouse*, colunar sobre DuckDB. A ampliação para novas fontes e categorias de consulta segue o mesmo padrão, o que reforça a transferibilidade da arquitetura.

Para a Informática na Educação, o EduQuery não é fonte primária de indicadores: cerca de 44% das alucinações *in-scope* ainda passam, e para usos críticos a revisão humana das fontes permanece indispensável. A camada de verificação determinística é transferível a qualquer sistema RAG educacional sobre dados estruturados. A separação entre conteúdo numérico e forma, adaptada por perfil, endereça um requisito recorrente no ensino e na divulgação científica, e a auditabilidade integral cumpre o requisito de reprodutibilidade priorizado pela trilha de Pesquisa em Informática na Educação. A arquitetura é ainda agnóstica ao domínio, bastando substituir os *marts Gold*.

6. Conclusão

A pergunta de pesquisa indagava em que medida uma camada de verificação determinística, interposta a um *pipeline* RAG sobre indicadores educacionais oficiais, aumenta a confiabilidade das respostas geradas por LLM, de forma reprodutível e independente do provedor. A evidência da Seção 4 responde afirmativamente: nos indicadores modelados, a camada intercepta 55,6% das alucinações que o RAG deixa passar e eleva a acurácia de 10% para 60%, com efeito grande e estatisticamente significativo; sustenta comportamento correto em 83,3% das consultas adversariais; e mantém-se em execução local. O ganho não decorre do RAG isoladamente, que sem verificação iguala a acurácia do LLM consultado diretamente.

Quatro frentes orientam a continuação: a ingestão dos microdados de PISA, TIMSS e PIRLS com valores plausíveis e pesos replicados [Mullis et al. 2023], para ampliar a interceptação; a verificação manual dos gabaritos e a repetição da bateria com $n \geq 3$ execuções; a avaliação com usuários reais nos três perfis (usabilidade, utilidade e confiança), pendente de aprovação ética; e mecanismos semânticos para o recorte adversarial.

Disponibilidade de Artefatos

O código-fonte do EduQuery e os *golden datasets* utilizados na avaliação descrita na Seção 4 estão disponíveis em repositório público, com as versões das dependências fixadas:

https://github.com/Thasys/analise_educacation_chatbot

O repositório contém as 84 consultas de avaliação com os respectivos valores esperados, tolerâncias e fontes primárias, nas três categorias da Subseção 4.1 (32 de valor único, 22 comparativas e 30 adversariais), além da orquestração multi-agente e da camada de *guardrails* determinísticos descritas na Seção 3.

Cuidados Éticos

Esta pesquisa não envolveu seres humanos: a avaliação empírica (Seção 4) foi conduzida sobre *golden datasets* construídos pelos autores a partir de fontes oficiais abertas. Avaliação futura com usuários reais será submetida a Comitê de Ética em Pesquisa.

Declaração sobre uso de Inteligência Artificial

Os autores utilizaram ferramentas de Inteligência Artificial generativa, baseadas em LLMs, durante o desenvolvimento deste trabalho em três frentes: (i) assistência na implementação do código do sistema descrito na Seção 3, incluindo refatorações e escrita de testes automatizados; (ii) assistência na elaboração dos *golden datasets* de avaliação (Subseção 4.1), com revisão humana subsequente; e (iii) revisão de redação em língua portuguesa do manuscrito. Todas as decisões metodológicas, as interpretações dos resultados quantitativos e as referências bibliográficas foram revisadas, validadas e, quando aplicável, corrigidas pelos autores. Nenhum dado empírico, valor numérico ou identificador de referência (DOI) foi obtido por geração de LLM sem validação contra a respectiva fonte primária.

Referências

- Angrist, N., Djankov, S., Goldberg, P. K., and Patrinos, H. A. (2021). Measuring human capital using global learning data. *Nature*, 592:403–408. DOI: <https://doi.org/10.1038/s41586-021-03323-7>.
- Armbrust, M., Ghodsi, A., Xin, R., and Zaharia, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. In *Conference on Innovative Data Systems Research (CIDR)*.
- Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*.
- Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). The goal question metric approach. In *Encyclopedia of Software Engineering*, pages 528–532. John Wiley & Sons.
- Carnoy, M., Khavenson, T., Fonseca, I., Costa, L., and Marotta, L. (2015). A educação brasileira está melhorando? evidências do PISA e do SAEB. *Cadernos de Pesquisa*, 45(157):450–485. DOI: <https://doi.org/10.1590/198053143331>.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. Disponível em: <https://arxiv.org/abs/2404.16130>. Acesso em: 14 jul. 2026.
- Fernandes, R. (2007). Índice de desenvolvimento da educação básica (IDEB). Série Documental 26, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

- Guardrails AI (2024). Guardrails AI: Adding guardrails to large language models. Disponível em: <https://github.com/guardrails-ai/guardrails>. Acesso em: 14 jul. 2026.
- Hanushek, E. A. and Woessmann, L. (2011). The economics of international differences in educational achievement. In Hanushek, E. A., Machin, S., and Woessmann, L., editors, *Handbook of the Economics of Education*, volume 3, pages 89–200. Elsevier.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12). DOI: <https://doi.org/10.1145/3571730>.
- LangChain (2024). LangGraph: Building stateful, multi-actor applications with LLMs. Disponível em: <https://github.com/langchain-ai/langgraph>. Acesso em: 14 jul. 2026.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.
- Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., and Wry, E. (2023). PIRLS 2021 international results in reading. Technical report, TIMSS & PIRLS International Study Center, Boston College.
- OECD (2024). Education at a glance 2024: OECD indicators. OECD Publishing, Paris. DOI: <https://doi.org/10.1787/c00cad36-en>.
- Raasveldt, M. and Mühleisen, H. (2019). DuckDB: an embeddable analytical database. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD)*, pages 1981–1984. DOI: <https://doi.org/10.1145/3299869.3320212>.
- Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., and Cohen, J. (2023). NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*.
- Schleicher, A. (2018). *World Class: How to Build a 21st-Century School System*. OECD Publishing, Paris. DOI: <https://doi.org/10.1787/9789264300002-en>.
- UNESCO (2020). Global education monitoring report 2020: Inclusion and education — all means all. UNESCO Publishing.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer, 2nd edition.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Proceedings of the First Conference on Language Modeling (COLM)*.
- Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H. (2024). Corrective retrieval augmented generation. Disponível em: <https://arxiv.org/abs/2401.15884>. Acesso em: 14 jul. 2026.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.