CATALOGAÇÃO DE QUESTÕES DE TI DA FGV: A INTELIGÊNCIA ARTIFICIAL DISPENSA A VALIDAÇÃO HUMANA?

Gabriel da Silva Narcisio[1], Francisco de Assis Pereira Vasconcelos de Arruda [2], Gislene Micarla Borges de Lima [1]

[1]UFERSA - Universidade Federal Rural do Semi-Árido, Angicos/RN, Brasil, gabriel.narcisio@alunos.ufersa.edu.br

Resumo: A Inteligência Artificial trouxe grandes inovações, mas também ampliou a dependência acrítica de tarefas, com usuários aceitando respostas sem validação. Este estudo examina cinco provas de Analista de Tecnologia da Informação da Fundação Getúlio Vargas (FGV) e verifica se a IA consegue catalogar questões com precisão semelhante à humana. Ao comparar a classificação manual com a realizada pelo modelo Gemini, observou-se que a tecnologia otimiza processos, mas não substitui a análise humana.

Palavras-chave: inteligência artificial; gemini; concursos; catalogação; análise

INTRODUÇÃO

Nos últimos anos, a Inteligência Artificial (IA) estabeleceu-se como uma ferramenta de grande impacto, apresentando um crescimento exponencial na educação e na análise de dados, trazendo inovações significativas, otimização de tempo e interesse pela evolução tecnológica. Para processar grandes volumes de informações, modelos de linguagem de grande escala (LLMs), como o ChatGPT e o Gemini da empresa Google, têm sido amplamente utilizados. Conforme apontam Cotton, Cotton e Shipway (2024), essas ferramentas generativas oferecem benefícios concretos, como maior eficiência e facilidade de acesso no desenvolvimento de atividades acadêmicas.

A pesquisa *Consumo e Uso da Inteligência Artificial no Brasil* revelou que 93% dos brasileiros utilizam alguma ferramenta de IA, ainda que 46% não compreendam o significado do termo (FUNDAÇÃO ITAÚ; DATAFOLHA, 2025). O levantamento ouviu 2.798 pessoas com 16 anos ou mais em todas as regiões do país, em julho de 2025. As aplicações mais comuns incluem redes sociais (89%), sistemas de recomendação (78%) e aplicativos de navegação (63%). Os dados evidenciam uma relação paradoxal: ampla penetração tecnológica no cotidiano, porém acompanhada de desconhecimento conceitual e uso predominantemente passivo (FUNDAÇÃO ITAÚ; DATAFOLHA, 2025).

A adoção acelerada dessas tecnologias tem gerado o problema da dependência excessiva. Lemos, Carvalho e Santos (2023) destacam que é cada vez mais comum que usuários deleguem tarefas às máquinas devido à grande praticidade, aceitando respostas sem a devida validação e correndo o risco de estagnar a própria capacidade de julgamento. Essa necessidade de mediação e monitoramento é uma realidade recorrente na análise de grandes exames educacionais. Conforme demonstrado por Lima *et al.* (2021), no contexto do Exame Nacional de Desempenho dos Estudantes (Enade), cujos procedimentos servem de base para metodologias de análise de conteúdo de provas, a aplicação de categorias estruturadas é indispensável para mapear áreas de conhecimentos predominantes e entender a organização de uma avaliação.

Essa mesma exigência analítica se estende ao cenário dos concursos públicos na área de Tecnologia da Informação (TI), onde a catalogação minuciosa de questões de provas constitui uma ferramenta essencial para analisar a forma como as bancas examinadoras cobram seus conteúdos programáticos. Embora a IA prometa aliviar a exaustiva leitura e classificação manual dessas provas, Cotton, Cotton e Shipway (2024) alertam que esses modelos podem gerar informações com erros factuais e inconsistências lógicas, tornando a validação humana um filtro crítico obrigatório.

No cenário dos concursos públicos brasileiros, existem diferentes bancas examinadoras responsáveis pela elaboração das provas, e cada uma delas pode apresentar características próprias na forma de cobrar os conteúdos previstos no edital. Em concursos voltados à área de TI, essa distribuição pode variar de acordo com o cargo ofertado, fazendo com que determinados conteúdos técnicos apareçam com maior ou menor frequência nas avaliações. Dessa forma, analisar provas de uma banca específica

permite observar recorrências na cobrança dos temas, identificar padrões na distribuição das questões e compreender melhor como os conhecimentos da área são organizados nas provas.

Diante dessa realidade, o presente estudo propõe e avalia um processo automatizado de catalogação de questões de concursos públicos da área de TI utilizando modelos de linguagem, em comparação à validação humana, investigando seus limites e potencialidades. Para isso, a pesquisa investigou cinco provas aplicadas pela Fundação Getúlio Vargas (FGV) para cargos na área de TI. A análise comparativa entre a classificação manual das questões e a gerada automaticamente pelo modelo Gemini busca demonstrar que, embora a tecnologia otimize processos, ainda não substitui a capacidade de validação humana.

MATERIAL E MÉTODOS

A presente pesquisa caracteriza-se como um estudo experimental de abordagem quantitativa e qualitativa. A escolha dos cadernos de questões da banca FGV justifica-se por sua relevância em concursos públicos nacionais e por sua adequação ao recorte deste estudo, voltado à análise de questões da área de TI. A utilização de provas elaboradas por uma mesma banca permite maior consistência na comparação dos resultados. A seleção das cinco provas para os concursos de: Analista Legislativo - Análise de Sistemas (ano 2022), DPE/RS - Desenvolvimento de Sistemas (ano 2023), Analista CVM - Sistemas e Desenvolvimento (ano 2024), SGB - Análise e Desenvolvimento de Sistemas (ano 2025) e AMAZUL - Desenvolvimento de Sistemas (ano 2025), buscou contemplar diferentes períodos e órgãos públicos com cargos relacionados ao campo de TI. A amostra está detalhada na Tabela 1, neste existe a especificação do tipo de prova que foi utilizada para cada concurso analisado.

Tabela 1. Relação das provas de TI analisadas.

Prova/Cargo	Ano	Tipo de Prova
Analista Legislativo - Análise de Sistemas	2022	Tipo 1 - Branca
DPE/RS - Desenvolvimento de Sistemas	2023	Tipo 1 - Branca
Analista CVM - Sistemas e Desenvolvimento	2024	Tipo 1 - Branca
SGB - Análise e Desenvolvimento de Sistemas	2025	Tipo 1 - Branca
AMAZUL - Desenvolvimento de Sistemas	2025	Tipo 1 - Branca

Para conduzir a análise dessa amostra, o ambiente de automação foi desenvolvido na linguagem de programação Java (Java Development Kit — JDK 17), utilizando o gerenciador de dependências Maven. A execução da pesquisa, com o objetivo de garantir uma comparação sistemática entre a avaliação humana e a inteligência artificial, foi estruturada em três etapas metodológicas consecutivas, conforme ilustrado na Figura 1.

1ª Etapa - Identificação e Criação

A primeira etapa da pesquisa consistiu na obtenção da amostra utilizada durante o estudo, realizando o *download* dos cadernos de provas da banca FGV (Bloco A). Após isso, foram selecionadas apenas as questões voltadas aos conhecimentos específicos da área de TI (Bloco B). Com a aplicação deste filtro, obteve-se um total de 205 questões objetivas para análise. Com as questões já separadas, foi realizada a criação de 11 rótulos categóricos referentes aos macroassuntos encontrados nas provas. Os rótulos foram definidos com base nos principais pilares temáticos presentes no conteúdo programático dos editais, com o objetivo de reduzir ambiguidades entre áreas conceitualmente próximas e garantir maior consistência no processo de catalogação. Entre eles estão: Desenvolvimento de Sistemas, Engenharia de *Software*, Arquitetura de *Software*, Banco de Dados, Infraestrutura e DevOps, Governança e Gestão de TI, Segurança da Informação, Inteligência Artificial e Business Intelligence, Governança de Dados, Sistemas Operacionais e Informática Básica. O mapeamento e a categorização de questões por áreas constituem uma prática metodológica consolidada para extração de indicadores em avaliações de TI, conforme evidenciado por Ramos, Oliveira e Braga (2024). Esses rótulos foram posteriormente utilizados na catalogação e organização das 205 questões analisadas (Bloco C).

2ª Etapa - Catalogação e Processamento

Esta etapa foi dedicada ao processo de extração e classificação das questões. Inicialmente, com o objetivo de estabelecer um parâmetro de comparação para a análise dos resultados, foi realizada uma

catalogação manual das questões com base nos rótulos previamente definidos (Bloco A). Esse procedimento foi conduzido pelo autor da pesquisa, seguindo critérios específicos para cada categoria temática. Dessa forma, a catalogação manual constituiu a linha de base utilizada para a comparação com as classificações produzidas pela IA ao longo do estudo. Nesse sentido, os casos em que a classificação da IA não coincidiu com a catalogação manual foram tratados como divergências de classificação, considerando a catalogação humana como parâmetro metodológico adotado para a análise comparativa. Paralelamente, para realizar a extração bruta dos textos presentes nos arquivos em *Portable Document Format* (PDF), utilizou-se a biblioteca Apache PDFBox (versão 2.0.30) (Bloco B). Já a classificação utilizando inteligência artificial foi conduzida por meio da biblioteca google-genai (versão 1.55.0), integrada à Interface de Programação de Aplicações (API) do modelo Gemini 3.5 Flash (versão de 19 de maio de 2026) (Bloco C). Visando garantir maior transparência metodológica, tanto o prompt utilizado durante os testes, quanto os arquivos em formato Comma-Separated Values (CSV) gerados pelo modelo encontram-se disponíveis em repositório público[1]. Além disso, buscando evitar sobrecarga durante a comunicação com a API e garantir maior estabilidade no processamento das requisições, foi implementado um intervalo de 15 segundos entre a análise de cada caderno de prova. Também foi definida a temperatura em 0,1 para manter o mínimo de criatividade do modelo escolhido. Como resultado final desta etapa, o modelo retornou os dados extraídos e organizados em formato CSV (Bloco D).

3ª Etapa - Análises

A etapa final consistiu no cruzamento entre os dados obtidos pela catalogação manual e os resultados gerados pela inteligência artificial (Bloco A). A partir dessa comparação, realizou-se uma análise quantitativa, buscando medir a taxa de concordância e de acerto do modelo em relação à linha de base manual adotada no estudo (Bloco B). Em seguida, também foi desenvolvida uma análise qualitativa, voltada à identificação de divergências de contexto e possíveis inconsistências da IA durante a interpretação das questões da banca (Bloco C).

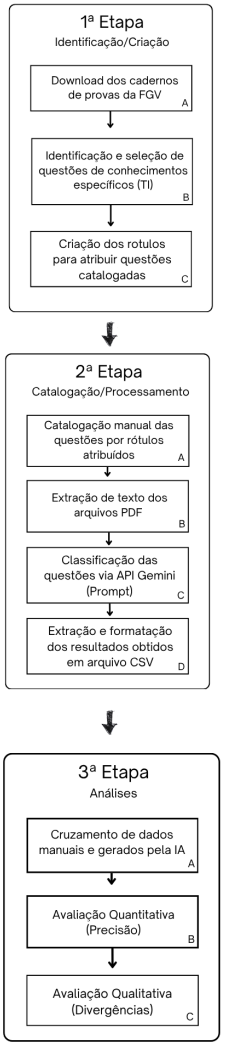


Figura 1. Fluxograma das etapas metodológicas.
Fonte: Elaborado pelo autor, adaptado de Lima et al. (2021).

RESULTADOS E DISCUSSÃO

No decorrer da experiência analítica, diversos aspectos chamaram a atenção no que diz respeito ao cruzamento entre o julgamento humano e a aplicação de técnicas de automação. Nesse contexto, observa-se que o uso de modelos de Inteligência Artificial em tarefas de classificação não se limita apenas à identificação mecânica de padrões, mas envolve também a aplicação de regras previamente definidas a partir dos dados analisados.

[1] Repositório do projeto contendo os dados complementares da pesquisa. Disponível em: https://github.com/GabriellNarcisio/catalogacao-ti-fgv-gemini.

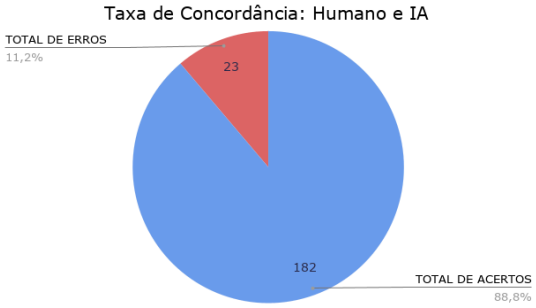


Gráfico 1. Taxa de concordância entre a catalogação humana e a Inteligência Artificial.

Em uma amostra total de 205 questões objetivas da área de TI, pertencentes à banca FGV, o modelo Gemini reproduziu corretamente o rótulo atribuído manualmente em 182 questões. Esse resultado, apresentado no Gráfico 1, corresponde a uma taxa de concordância de 88,8%. Embora esse percentual indique um desempenho satisfatório da IA, sua interpretação exige cuidado, pois a taxa geral de concordância não permite identificar, de forma isolada, em quais provas ou macroassuntos ocorreram as maiores dificuldades de classificação.

Para compreender melhor esses resultados, os dados da planilha de catalogação foram analisados separadamente por prova e por macroassunto. Essa organização permitiu comparar a classificação manual com a classificação realizada pela IA, observando em quais casos o modelo apresentou maior ou menor concordância.

A partir dessa análise, verificou-se que o desempenho do modelo variou entre as provas analisadas. A maior concordância ocorreu na prova de Analista Legislativo, com 28 classificações concordantes em um total de 30 questões. Já a menor concordância foi observada na prova da DPE/RS, com 24 classificações concordantes em 30 questões. Esses resultados indicam que, mesmo com uma taxa geral elevada, a precisão da classificação pode variar de acordo com a organização temática de cada prova e com a proximidade entre os conteúdos cobrados.

Também foi analisada a distribuição dos macroassuntos presentes na catalogação manual. Conforme apresentado no Gráfico 2, o tema Desenvolvimento de Sistemas foi o mais frequente, concentrando 24,88% do total, o que corresponde a 51 questões. Em seguida, apareceram Infraestrutura e DevOps e Sistemas Operacionais, ambos com 27 questões. Por outro lado, o tema Inteligência Artificial e Business Intelligence apresentou menor representatividade, com apenas 5 questões.

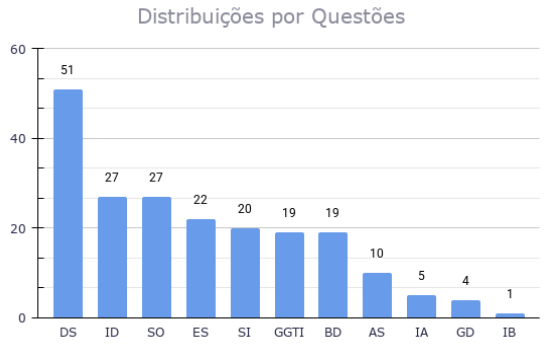


Gráfico 2. Distribuição percentual dos assuntos cobrados e catalogados manualmente

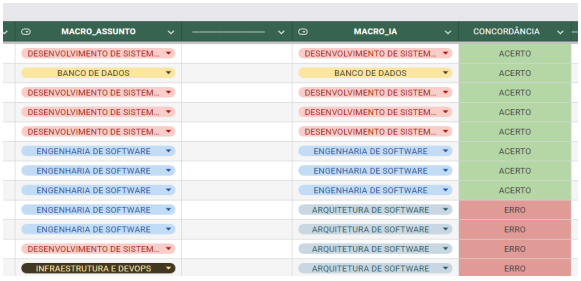
Figura 2. Abreviaturas. Fonte: autoria própria.

A diferença na quantidade de questões por macroassunto também reflete a composição das provas analisadas, uma vez que cada cargo pode priorizar conteúdos distintos dentro da área de TI. Assim, a distribuição temática observada não indica, necessariamente, um desequilíbrio metodológico, mas evidencia a diversidade de enfoques presentes nos cadernos selecionados. Por esse motivo, a taxa geral de concordância deve ser interpretada em conjunto com a organização temática da amostra.

Na análise qualitativa das 23 questões discordantes, observou-se que as divergências não ocorreram de forma aleatória. Elas apareceram principalmente em regiões de fronteira entre áreas conceitualmente próximas, nas quais a separação entre os conteúdos pode ser mais difícil. A principal diferença ocorreu entre Engenharia de Software e Arquitetura de Software, indicando uma tendência do modelo em associar determinados conteúdos classificados manualmente como Engenharia de Software ao rótulo Arquitetura de Software.

Uma possível explicação para esse resultado está na proximidade entre essas áreas. Conceitos como modelagem de sistemas, Linguagem de Modelagem

Unificada (UML), padrões arquiteturais e processos de desenvolvimento de software podem aparecer de forma integrada nas questões analisadas. Isso pode dificultar a separação exata entre os rótulos utilizados na catalogação. Essa situação é apresentada na Figura 3, que mostra um exemplo de divergência entre a catalogação manual e a catalogação realizada pela IA em uma questão situada em uma área de sobreposição conceitual.



MACRO_ASSUNTO		MACRO_IA	CONCORDÂNCIA
DESENVOLVIMENTO DE SISTEM...		DESENVOLVIMENTO DE SISTEM...	ACERTO
BANCO DE DADOS		BANCO DE DADOS	ACERTO
DESENVOLVIMENTO DE SISTEM...		DESENVOLVIMENTO DE SISTEM...	ACERTO
DESENVOLVIMENTO DE SISTEM...		DESENVOLVIMENTO DE SISTEM...	ACERTO
DESENVOLVIMENTO DE SISTEM...		DESENVOLVIMENTO DE SISTEM...	ACERTO
ENGENHARIA DE SOFTWARE		ENGENHARIA DE SOFTWARE	ACERTO
ENGENHARIA DE SOFTWARE		ENGENHARIA DE SOFTWARE	ACERTO
ENGENHARIA DE SOFTWARE		ENGENHARIA DE SOFTWARE	ACERTO
ENGENHARIA DE SOFTWARE		ARQUITETURA DE SOFTWARE	ERRO
ENGENHARIA DE SOFTWARE		ARQUITETURA DE SOFTWARE	ERRO
DESENVOLVIMENTO DE SISTEM...		ARQUITETURA DE SOFTWARE	ERRO
INFRAESTRUTURA E DEVOPS		ARQUITETURA DE SOFTWARE	ERRO

Figura 3. Divergência entre a catalogação manual e a catalogação da IA em área de sobreposição conceitual. Fonte: autoria própria.

Esse comportamento pode ser compreendido considerando que modelos de classificação são treinados para produzir uma única saída categórica a partir das informações de entrada, o que leva à simplificação de nuances e proximidades semânticas entre as classes em uma decisão final.

Dessa forma, os resultados reforçam a importância da mediação humana no processo de validação das classificações realizadas por sistemas automatizados. Mais do que observar o elevado índice de concordância alcançado, é necessário compreender as limitações do modelo em contextos nos quais as categorias apresentam proximidade semântica.

Além disso, os resultados indicam que a IA pode atuar como apoio na triagem inicial das questões, reduzindo o esforço de leitura e organização dos dados. No entanto, as divergências observadas mostram que a revisão humana continua necessária, principalmente quando os conteúdos avaliados apresentam proximidade conceitual. Dessa forma, o uso combinado entre automação e validação humana mostra-se mais adequado, pois permite manter a agilidade do processo sem comprometer o rigor da catalogação.

CONCLUSÃO

Este estudo evidenciou que a utilização da Inteligência Artificial para a catalogação de questões de concursos de TI apresenta potencial significativo, embora ainda enfrente desafios em áreas de fronteira conceitual. Observou-se que o modelo Gemini alcançou uma expressiva taxa de concordância de 88,8%, contribuindo para a otimização do processo de triagem. No entanto, o papel humano permanece fundamental, indo além da simples operação do sistema, especialmente em situações que exigem maior interpretação crítica, como ocorreu nos casos de proximidade entre Engenharia e Arquitetura de Software.

Os autores deste trabalho reconhecem a existência de limitações quanto à generalização dos resultados, principalmente em razão do escopo restrito a cinco provas da banca FGV. Ainda assim, essa limitação reforça a importância de aprofundamentos metodológicos em pesquisas futuras. Como próximos passos, pretende-se ampliar a quantidade de provas analisadas, incluir outras bancas examinadoras da mesma área e integrar os resultados obtidos à modelagem estrutural do projeto, buscando aprimorar os processos de extração e classificação automatizada em arquivos PDF.

Em suma, na perspectiva dos autores, a experiência demonstrou que modelos de linguagem não atuam como substitutos absolutos da análise crítica humana, mas sim como ferramentas de apoio estratégico. Nesse contexto, a validação humana continua desempenhando papel essencial para garantir maior rigor e confiabilidade em processos avaliativos e educacionais.

AGRADECIMENTOS

Ao Programa de Educação Tutorial (PET) pelo apoio concedido, aos professores pelo direcionamento acadêmico e aos colegas que contribuíram ativamente na testagem e validação do código, viabilizando o desenvolvimento deste estudo e a elaboração deste trabalho.

REFERÊNCIAS

COTTON, D. R. E.; COTTON, P. A.; SHIPWAY, J. R. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. **Innovations in Education and Teaching International**, v. 61, n. 2, p. 228-239, 2024.

FUNDAÇÃO ITAÚ; DATAFOLHA. **Consumo e uso da Inteligência Artificial no Brasil**. São Paulo: Observatório Fundação Itaú, 2025. Disponível em: https://www.fundacaoitau.org.br/noticias/noticias/93-das-pessoas-utilizam-ferramenta-de-ia-diz-pesquisa-da-fundacao-itau-e-datafolha. Acesso em: 13 jun. 2026.

LEMOS, G. M.; CARVALHO, G. A.; SANTOS, J. V. C. Inteligência artificial: riscos, benefícios e uso responsável. **Apoena**, v. 6, p. 517-527, 2023.

LIMA, P. S. N.; AMBRÓSIO, A. P. L.; OLIVEIRA, J. L. S.; CARVALHO, C. L. Análise de conteúdo das provas do Enade para os alunos do curso de Bacharelado em Ciência da Computação. **Revista Brasileira de Informática na Educação (RBIE)**, v. 29, p. 385-413, 2021.

RAMOS, M. S.; OLIVEIRA, P. T. G.; BRAGA, A. R. Análise do Desempenho dos Estudantes do Curso Redes de Computadores da Universidade Federal do Ceará no ENADE. In: WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 32., 2024, Brasília/DF. **Anais do XXXII Workshop sobre Educação em Computação (WEI 2024)**. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 658-668.