Avaliação do Uso de Grandes Modelos de Linguagem na Análise de Riscos em Segurança da Informação: Um Estudo Comparativo com Especialistas Humanos

Gustavo Roberto Pinto

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2026

Gustavo Roberto Pinto

Avaliação do Uso de Grandes Modelos de Linguagem na Análise de Riscos em Segurança da Informação: Um Estudo Comparativo com Especialistas Humanos

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Rodrigo Sanches Miani

Uberlândia
2026

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

P659 Pinto, Gustavo Roberto, 1982-
2026 Avaliação do Uso de Grandes Modelos de Linguagem na Análise de Riscos em Segurança da Informação: Um Estudo Comparativo com Especialistas Humanos [recurso eletrônico] / Gustavo Roberto Pinto. - 2026.

Orientador: Rodrigo Sanches Miani.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2026.551
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Miani, Rodrigo Sanches,1983-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 15/2026, PPGCO				
Data:	08 de Julho de 2026	Hora de início:	09:00	Hora de encerramento:	11:35
Matrícula do Discente:	12422CCP013				
Nome do Discente:	Gustavo Roberto Pinto				
Título do Trabalho:	Avaliação do Uso de Grandes Modelos de Linguagem na Análise de Riscos em Segurança da Informação: Um Estudo Comparativo com Especialistas Humanos				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Sistemas de Computação				
Projeto de Pesquisa de vinculação:	----------				

Reuniu-se por presencialmente, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Diego Nunes Molinos - FACOM/UFU, Bruno Bogaz Zarpelão - CCE/UEL e Rodrigo Sanches Miani - FACOM/UFU, orientadora do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Todos os membros da banca e o aluno(a) participaram da cidade de Uberlândia.

Iniciando os trabalhos o(a) presidente da mesa, Prof. Dr. Rodrigo Sanches Miani , apresentou a Comissão Examinadora e o(a) candidato(a), agradeceu a presença do público, e concedeu ao(à) Discente a palavra para a exposição do seu trabalho.

A seguir o senhora presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente

ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Bruno Bogaz Zarpelão**, **Usuário Externo**, em 10/07/2026, às 11:42, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rodrigo Sanches Miani**, **Professor(a) do Magistério Superior**, em 10/07/2026, às 12:00, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Diego Nunes Molinos**, **Professor(a) do Magistério Superior**, em 10/07/2026, às 14:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **7455715** e o código CRC **4161A9C1**.

Referência: Processo nº 23117.044151/2026-13 SEI nº 7455715

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada **"Avaliação do Uso de Grandes Modelos de Linguagem na Análise de Riscos em Segurança da Informação: Um Estudo Comparativo com Especialistas Humanos"** por **Gustavo Roberto Pinto** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação.**

Uberlândia, 08 de Julho de 2026

Orientador:_____

Prof. Dr. Rodrigo Sanches Miani

Universidade Federal de Uberlândia

Banca Examinadora:

Membro da banca 1:_____

Prof. Dr. Diego Nunes Molinos

Universidade Federal de Uberlândia

Membro da banca 2:_____

Prof. Dr. Bruno Bogaz Zarpelão

Universidade Estadual de Londrina

Dedico este trabalho a Deus e à minha família, fonte de amor, força e inspiração em todos os momentos. Aos meus pais e irmãos, por caminharem ao meu lado ao longo da vida. À minha esposa Loren, às nossas companheirinhas Eva e Meg, e ao nosso pequeno anjo que viverá para sempre em nossos corações.

Agradecimentos

Agradeço primeiramente a Deus, por me conceder saúde, força, sabedoria e perseverança para superar os desafios e receber as glórias desta existência. Aos meus pais Sandra e Afrânio, pelo amor incondicional, pelos ensinamentos, pelos valores transmitidos ao longo da vida e por sempre acreditarem em mim, mesmo nos momentos mais difíceis. Aos meus irmãos José e Afrânio, pelo companheirismo, amizade e apoio constante. À minha esposa Loren, por caminhar ao meu lado todos os dias, compartilhando sonhos, desafios e conquistas. Obrigado pelo amor, pela compreensão nos momentos de ausência e por construir comigo a família que tanto amo. Ao meu orientador, Prof. Dr. Rodrigo Miani, pela orientação, confiança, disponibilidade e imenso apoio durante o desenvolvimento desta pesquisa. Seus conhecimentos e contribuições foram fundamentais para a realização deste trabalho. Ao meu amigo, colega de trabalho e companheiro de mestrado, Arthur Labaki, pelas discussões, trocas de ideias, incentivos e aprendizados compartilhados ao longo dessa jornada acadêmica e profissional. Aos meus sócios Henrique e Afrânio, por acreditarem e construírem comigo o sonho da EVA CYBERSECURITY. Em especial ao Afrânio, meu irmão, amigo e parceiro de jornada, que esteve ao meu lado desde o início dessa caminhada. Compartilhamos desafios, incertezas, conquistas e aprendizados que ultrapassam os limites da vida profissional. A todos os colaboradores da EVA CYBERSECURITY, pela dedicação, pelo ambiente de aprendizado, colaboração e constante incentivo ao desenvolvimento profissional e acadêmico. O apoio e a convivência diária contribuíram significativamente para minha formação e para a realização desta pesquisa. Aos nossos clientes, pela confiança, parceria e pelos desafios compartilhados ao longo dos anos. Cada projeto, conversa e problema enfrentado em conjunto contribuiu para meu crescimento profissional e ajudou a construir a experiência prática que motivou e enriqueceu esta pesquisa. Agradeço também aos profissionais que dedicaram seu tempo para participar voluntariamente dos experimentos desta dissertação. Sem a contribuição de cada um deles, este trabalho não teria sido possível. Por fim, agradeço a todos que, de alguma forma, contribuíram para esta conquista e fizeram parte desta caminhada. Cada apoio recebido, cada ensinamento compartilhado e cada palavra de incentivo tiveram im-

portância nesta trajetória. Esta conquista é resultado não apenas do meu esforço, mas também da presença e contribuição de todas as pessoas que caminharam ao meu lado ao longo desses anos.

"No que diz respeito ao empenho, ao compromisso, ao esforço e à dedicação, não existe meio termo. Ou você faz uma coisa bem feita ou não faz."

Ayrton Senna, do Brasil

Resumo

A crescente sofisticação das ameaças cibernéticas e a constante evolução do cenário tecnológico impõem desafios significativos às organizações, especialmente diante da escassez global de profissionais qualificados em segurança da informação. Nesse contexto, a Inteligência Artificial Generativa (GenAI) e os Grandes Modelos de Linguagem (LLM) surgem como alternativas promissoras para apoiar atividades relacionadas à análise e avaliação de riscos de segurança da informação. Apesar de seu potencial, os LLMs também apresentam limitações, como a geração de informações imprecisas ou alucinatórias, capazes de influenciar negativamente processos decisórios. Esta dissertação investiga o comportamento dos LLMs na avaliação de cenários de risco de segurança da informação em comparação com especialistas humanos, destacando a importância da supervisão humana e da integração entre humanos e máquinas nesse processo. Para isso, realizamos uma pesquisa contendo cenários de risco com participantes humanos, cujas respostas foram comparadas às respostas produzidas por cinco LLMs populares. Os resultados sugerem que os LLMs tenderam a subestimar os riscos de segurança em comparação aos participantes humanos. Assim, embora possam ser úteis como ferramentas de apoio, seu uso como avaliadores autônomos em processos de avaliação de risco de segurança da informação deve ser considerado com cautela.

Palavras-chave: Inteligência Artificial, Segurança de Computadores, Fatores Humanos, Análise de Riscos.

Abstract

The increasing sophistication of cyber threats and the continuous evolution of the technological landscape impose significant challenges on organizations, especially given the global shortage of qualified cybersecurity professionals. In this context, Generative Artificial Intelligence and Large Language Models emerge as promising alternatives to support cybersecurity risk analysis and assessment activities. Despite their potential, Large Language Models also present limitations, such as generating inaccurate or 'hallucinated' information capable of negatively influencing decision-making processes. This thesis investigates the behavior of Large Language Models in cybersecurity risk assessment scenarios compared to human experts, highlighting the importance of human oversight and the integration of humans and machines in this process. To achieve this, a questionnaire containing risk scenarios was applied to human participants, whose responses were compared with those generated by five popular Large Language Models. The results suggest that LLMs tended to underestimate security risks compared to human participants. Therefore, although they may be useful as support tools, their use as autonomous evaluators in information security risk assessment processes should be considered with caution.

Keywords: Artificial intelligence, Computer security, Human factors, Risk analysis.

Lista de ilustrações

Lista de tabelas

Lista de siglas

BERT Bidirectional Encoder Representations from Transformers

CIS *Critical Security Controls*

CSF *NIST Cybersecurity Framework*

CWE *Common Weakness Enumeration*

CVE *Common Vulnerabilities and Exposures*

MITRE ATT e CK *Adversarial Tactics, Techniques, and Common Knowledge*

CISA *Cybersecurity and Infrastructure Security Agency*

CNN *Convolutional Neural Networks*

COBIT *Control Objectives for Information and Related Technology*

CET *Cybersecurity Evaluation Tool*

FAIR *Factor Analysis of Information Risk*

GenAI Inteligência Artificial Generativa

GAN *Generative Adversarial Networks*

GRC Governança, Riscos e Conformidade

HITL Human-in-the-loop

ISO/IEC 27001 Norma para Sistemas de Gestão de Segurança da Informação

ISO/IEC 31000 Norma para Gestão de Riscos

ISO/IEC 42001 Norma para Gestão de Inteligência Artificial

IA Inteligência Artificial

ICS Sistemas de controle industrial

IOT *Internet of Things*

IIoT *Industrial Internet of Things*

LLM Grandes Modelos de Linguagem

MAE Mean Absolute Error

NIST *National Institute of Standards and Technology*

NIST/SP 800-30 *Guide for Conducting Risk Assessments*

PME Pequenas e médias empresas

PPFLE *Privacy-Preserving Fixed-Length Encoding*

RAG *Retrieval-Augmented Generation*

RNN *Recurrent Neural Networks*

RLHF *Reinforcement Learning from Human Feedback*

RFC *Requests For Comments*

VAE *Variational Autoencoder*

xAI *Explainable Artificial Intelligence*

Sumário

Capítulo 1

Introdução

O aumento da complexidade das ameaças cibernéticas tem ampliado os desafios enfrentados pelas organizações na proteção de seus ativos, processos e informações. Além dos impactos econômicos associados ao cibercrime, ataques podem explorar vulnerabilidades conhecidas ou de dia zero, comprometer cadeias de suprimentos, empregar técnicas de engenharia social e envolver ameaças persistentes avançadas. Esse cenário, marcado pela automação e pela crescente sofisticação dos agentes maliciosos, reforça a necessidade de abordagens estruturadas para identificar, analisar e tratar riscos de segurança da informação (World Economic Forum, 2025b; SentinelOne, 2026; WINDER, 2026). A gestão desses riscos depende da seleção, implementação, manutenção e monitoramento de controles técnicos e não técnicos compatíveis com o contexto, os ativos, as ameaças e as necessidades de cada organização. Quando aplicados de forma proporcional aos riscos identificados, esses controles contribuem para reduzir vulnerabilidades, limitar impactos e fortalecer a resiliência operacional e a governança. Entretanto, a análise e a priorização dos riscos exigem conhecimento especializado, interpretação contextual e capacidade de tomada de decisão em ambientes caracterizados por incerteza (NIST, 2012).

Esse desafio é ampliado pela escassez global de profissionais qualificados em cibersegurança. Estimativas apontam uma lacuna de aproximadamente 4,7 milhões de profissionais, enquanto parcela significativa das organizações identifica a falta de competências como uma barreira à resiliência cibernética (FORUM, 2024; World Economic Forum, 2025a). A evolução acelerada das ameaças e dos ambientes tecnológicos aumenta a demanda por profissionais com conhecimentos atualizados, ao mesmo tempo em que dificulta a formação, a atração e a retenção de especialistas capazes de apoiar processos complexos de avaliação e tratamento de riscos ((ISC)², 2024).

Nesse contexto, a GenAI, especialmente por meio dos LLMs, tem despertado interesse crescente como ferramenta de apoio a atividades que demandam análise de informações, produção de conteúdo e suporte à tomada de decisão. Na área de segurança da informação, esses modelos vêm sendo empregados em tarefas como análise de vulnerabilidades, resposta a incidentes, geração de documentação, automação de atividades operacionais

e apoio à avaliação de riscos, indicando potencial para auxiliar profissionais diante da crescente demanda por especialistas (YAO et al., 2024).

Entretanto, apesar dos avanços observados, os LLMs apresentam limitações importantes. Por serem modelos probabilísticos treinados para prever sequências de texto, podem produzir respostas inconsistentes, imprecisas ou sem fundamentação adequada, especialmente em tarefas que exigem julgamento especializado e compreensão contextual. Essas limitações reforçam a necessidade de avaliar, de forma sistemática, até que ponto esses modelos podem apoiar processos críticos de segurança da informação e em quais situações a supervisão humana permanece indispensável (PANKAJAKSHAN et al., 2024). Embora a literatura apresente avanços significativos na aplicação de LLMs à cibersegurança, observa-se uma lacuna em estudos que comparem diretamente as avaliações produzidas por diferentes modelos de linguagem com aquelas realizadas por profissionais humanos diante dos mesmos cenários estruturados de análise de riscos. Essa lacuna torna-se ainda mais evidente quando se considera a influência do nível de experiência dos participantes sobre as avaliações realizadas, aspecto pouco explorado pelos trabalhos existentes.

Diante dessa lacuna, esta dissertação investiga em que medida Grandes Modelos de Linguagem são capazes de produzir avaliações de risco alinhadas às realizadas por profissionais de segurança da informação quando submetidos aos mesmos cenários estruturados. Para isso, são comparadas as avaliações produzidas por cinco LLMs amplamente utilizados e por profissionais com diferentes níveis de experiência, buscando identificar o grau de alinhamento entre esses grupos, bem como eventuais diferenças sistemáticas na atribuição dos níveis de risco. Ao produzir evidências sobre o comportamento dos LLMs nesse contexto, esta pesquisa busca contribuir para a compreensão de suas potencialidades e limitações como ferramentas de apoio à avaliação de riscos em segurança da informação, fornecendo subsídios para seu uso mais criterioso em processos de tomada de decisão e para o desenvolvimento de pesquisas futuras sobre colaboração entre especialistas humanos e inteligência artificial.

1.1 Objetivos e Desafios da Pesquisa

Embora Grandes Modelos de Linguagem tenham demonstrado potencial para apoiar diversas atividades relacionadas à segurança da informação, ainda não está suficientemente esclarecido em que medida suas avaliações de risco se alinham às produzidas por profissionais humanos quando submetidos aos mesmos cenários estruturados. Também permanece pouco compreendido como esse alinhamento varia em função do nível de experiência dos profissionais e quais são as limitações desses modelos para apoiar ou substituir avaliações realizadas por especialistas (YAO et al., 2024; GENNARI et al., 2024; FERRAG et al., 2024b; PANKAJAKSHAN et al., 2024; ZHAO et al., 2023).

Nesse contexto, esta pesquisa busca investigar o grau de alinhamento entre as avalia-

ções produzidas por cinco Grandes Modelos de Linguagem (*ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro*) e aquelas realizadas por profissionais de segurança da informação, utilizando cenários padronizados baseados nos dezoito controles do *Critical Security Controls* (CIS) (Center for Internet Security, 2021). No escopo deste trabalho, a análise não busca estabelecer um valor verdadeiro absoluto para cada cenário, mas comparar as avaliações dos modelos com o julgamento de profissionais da área, utilizado como referência empírica para analisar o comportamento dos LLMs em processos de avaliação de riscos, buscando esclarecer em que medida os LLMs podem ser utilizados com — ou sem — a supervisão de humanos na avaliação de riscos.

1.1.1 Objetivos específicos

- ❏ Criar um modelo de **referência** (*Benchmark*), para análise de riscos usando LLMs. O modelo será criado com 18 cenários de riscos de segurança da informação, usando como referência o CIS, para ser respondido por grandes modelos de linguagem amplamente utilizados e especialistas humanos nas áreas de tecnologia e segurança, que classificarão qual o grau de risco de cada cenário proposto;

- ❏ Criar um método preliminar de **calibragem** para avaliar como os LLMs se comportam ao atribuir um grau de risco a variações do mesmo cenário base de risco de segurança da informação, com o intuito de validar o comportamento dos modelos quanto à habilidade na avaliação de riscos, melhor explicado na Subseção 4.2.2;

- ❏ Criar um **conjunto de dados com as respostas dos 5 principais LLMs** respondendo aos 18 cenários de risco de segurança da informação, do cenário base;

- ❏ Comparar as respostas dos 5 principais modelos com as respostas dos participantes humanos e gerar um **Benchmark com o desempenho dos modelos** quanto à atribuição de riscos de segurança em relação aos participantes humanos.

1.2 Perguntas de pesquisa

As perguntas de pesquisa definidas nesta dissertação modelam a análise comparativa entre os LLMs e profissionais humanos na avaliação de riscos de segurança da informação, considerando tanto a confiabilidade dos modelos quanto o desempenho humano em diferentes níveis de experiência profissional.

Pergunta de Pesquisa 1 (*PP1*):

Em que medida os LLMs podem ser considerados confiáveis para realizar avaliações de risco em segurança da informação sem a participação de especialistas humanos?

Pergunta de Pesquisa 2 (*PP2*):

Como os LLMs utilizados nesta dissertação se comparam aos profissionais humanos na avaliação de riscos nos 18 cenários?

- ❏ PP2.1 Profissionais Seniores e Especialistas: Como o desempenho dos LLMs se alinha ou difere daquele observado entre os profissionais que possuem maior experiência e conhecimento técnico?
- ❏ PP2.2 Outros Profissionais (Não Seniores/Especialistas): Como os LLMs se comparam aos profissionais de outros níveis (ex.: júnior, pleno e estagiário) no que diz respeito à precisão e consistência na atribuição de níveis de risco?

1.3 Contribuições

Esta pesquisa amplia a literatura sobre a aplicação de Grandes Modelos de Linguagem em segurança da informação ao comparar, de forma sistemática, avaliações produzidas por diferentes LLMs e por profissionais humanos utilizando cenários estruturados baseados no CIS. Diferentemente de trabalhos focados apenas em tarefas factuais, classificação automática ou detecção de ameaças, esta pesquisa analisa como modelos de linguagem se comportam em processos de avaliação de risco que envolvem percepção, experiência profissional, subjetividade e tomada de decisão em contextos organizacionais. Além disso, esta dissertação propõe uma abordagem experimental baseada em cenários estruturados fundamentados nos controles CIS, permitindo uma avaliação comparável entre julgamentos humanos e inferências produzidas por modelos de linguagem. Assim, esta dissertação propõe as seguintes contribuições:

- ❏ Proposição de uma abordagem experimental para avaliar o uso de Grandes Modelos de Linguagem na análise de riscos em segurança da informação;
- ❏ Construção de um conjunto estruturado de cenários de risco baseados nos 18 controles do *Center for Internet Security* (CIS), permitindo avaliações padronizadas e comparáveis;
- ❏ Realização de uma comparação metodológica entre avaliações produzidas por especialistas humanos e por cinco LLMs amplamente utilizados;
- ❏ Investigação do impacto do nível de experiência profissional humana na percepção e classificação de riscos;
- ❏ Identificação de padrões de subestimação de riscos pelos LLMs em relação aos participantes humanos;

- ❑ Evidências experimentais que reforçam a necessidade de abordagens híbridas Human-in-the-loop (HITL) em processos de análise de risco apoiados por Inteligência Artificial.

1.4 Organização da Dissertação

Esta dissertação está estruturada em seis capítulos, organizados de forma a apresentar progressivamente os fundamentos teóricos, a contextualização do problema, a metodologia adotada, os experimentos conduzidos e as conclusões obtidas a partir da pesquisa.

O Capítulo 1 apresenta a introdução da dissertação, contextualizando os desafios contemporâneos relacionados à segurança da informação, ao aumento da complexidade das ameaças digitais e à escassez global de profissionais qualificados em cibersegurança. O capítulo também discute o potencial da Inteligência Artificial Generativa e dos LLMs como ferramentas de apoio à análise de riscos, destacando seus benefícios, limitações e riscos associados. Além disso, são apresentados os objetivos da pesquisa e as perguntas de pesquisa abordadas nesta dissertação.

O Capítulo 2 apresenta a fundamentação teórica necessária para sustentar a pesquisa. Inicialmente, são discutidos os conceitos fundamentais de Inteligência Artificial (IA), GenAI e LLMs, incluindo suas capacidades, aplicações e desafios éticos. Na Seção 2.2, o capítulo aborda conceitos relacionados à gestão de riscos em segurança da informação, *frameworks* amplamente utilizados na área e aspectos associados ao julgamento humano na análise de riscos, incluindo fatores cognitivos, experiência profissional e tomada de decisão em cenários de incerteza.

O Capítulo 3 apresenta os trabalhos relacionados e discute pesquisas recentes na interseção entre Inteligência Artificial e segurança da informação. O capítulo organiza a literatura em diferentes grupos, incluindo Benchmarks para avaliação de capacidades de LLMs, modelos especializados para tarefas de segurança e *frameworks* voltados à análise de riscos organizacionais.

O Capítulo 4 descreve a metodologia utilizada na condução da pesquisa. O método proposto é estruturado em quatro etapas principais: criação de dados, testes com LLMs, pesquisa com participantes humanos e análise dos resultados. O capítulo detalha a construção dos cenários baseados nos controles do CIS, a definição dos questionários aplicados, o processo de recrutamento dos participantes humanos, a caracterização da amostra e os procedimentos adotados para avaliação e comparação dos resultados obtidos pelos modelos de linguagem e pelos participantes humanos.

O Capítulo 5 apresenta os experimentos realizados e a análise dos resultados obtidos ao longo da pesquisa. São apresentados os dados demográficos dos participantes humanos, as análises relacionadas às perguntas de pesquisa propostas e as comparações entre os julgamentos realizados pelos LLMs e pelos profissionais humanos em diferentes níveis de

experiência. O capítulo também discute os padrões identificados nos resultados, incluindo tendências de subestimação de riscos pelos modelos, níveis de correlação entre humanos e IA e as limitações observadas durante os experimentos.

Por fim, o Capítulo 6 apresenta as conclusões desta dissertação, destacando as principais contribuições científicas e práticas do trabalho. O capítulo discute os limites atuais da utilização de LLMs em processos de análise de riscos de segurança da informação, reforçando a importância da integração entre especialistas humanos e sistemas baseados em Inteligência Artificial. Além disso, são apresentadas propostas de trabalhos futuros, incluindo a ampliação da base experimental, o uso de técnicas avançadas de engenharia de *Prompts*, abordagens híbridas Humano-IA e o treinamento de modelos especializados em segurança da informação.

Capítulo 2

Fundamentação teórica

O Capítulo 2 desta dissertação tem como objetivo estabelecer as bases conceituais necessárias para a compreensão da relação entre grandes modelos de linguagem e o julgamento humano na avaliação de riscos em segurança da informação. Na Seção 2.1 são apresentados os conceitos fundamentais de Inteligência Artificial, com ênfase na GenAI e nos LLMs, destacando suas capacidades, aplicações e limitações. Na Seção 2.2, são abordados os fundamentos da gestão de riscos em segurança da informação, incluindo *frameworks* amplamente adotados, conceitos clássicos de análise de risco e as particularidades desse processo, como, por exemplo, os aspectos cognitivos envolvidos na avaliação de riscos, explorando como o julgamento humano é influenciado por heurísticas, vieses cognitivos e fatores emocionais, conforme descrito na literatura de tomada de decisão sob incerteza.

Também são incorporadas perspectivas relacionadas ao desenvolvimento de expertise, destacando como a experiência e a prática deliberada influenciam a consistência e a acurácia das avaliações de risco. Essa abordagem permite explicar não apenas a presença de vieses, mas também a variabilidade observada entre profissionais com diferentes níveis de experiência na análise de riscos. Por fim, a Seção 2.3 conclui a fundamentação da investigação proposta neste trabalho, que busca analisar o desempenho de LLMs em comparação com participantes humanos, bem como explorar o potencial de integração entre capacidades humanas e artificiais no suporte à tomada de decisão em cenários de risco.

2.1 Inteligência artificial

Esta Seção apresenta os conceitos fundamentais de Inteligência Artificial, com ênfase na Inteligência Artificial Generativa e nos Grandes Modelos de Linguagem. São discutidos seus princípios de funcionamento, principais aplicações e limitações, com foco em seu potencial como ferramentas de suporte à tomada de decisão em contextos complexos, como a segurança da informação.

2.1.1 Inteligência Artificial Generativa

A Inteligência Artificial Generativa (GenAI) e os Grandes Modelos de Linguagem (LLMs) demonstrados na Tabela 1, como o ChatGPT da OpenAI, Gemini do Google, Llama da Meta e Claude da Anthropic, ganharam destaque nos últimos anos por sua capacidade de auxiliar profissionais em diversas áreas e tarefas. Esses modelos utilizam vastas redes neurais e técnicas de aprendizado profundo para processar e gerar texto, código e até mesmo imagens de forma natural, imitando funções cognitivas humanas complexas, como resolução de problemas, criatividade e tomada de decisões (ZHAO et al., 2023; NAVEED et al., 2023).

Tabela 1 – LLMs investigados nesta dissertação

Modelo	**Org.**	**Base**	**Características**	**Base Bibliográfica**
ChatGPT 5	OpenAI	Transformer + RLHF	Geração avançada de linguagem e suporte a múltiplas tarefas	(BROWN et al., 2020; OpenAI, 2023; OUYANG et al., 2022)
Llama 4 Scout	Meta	Transformer	Modelos eficientes e adaptáveis para pesquisa	(TOUVRON et al., 2023a; TOUVRON et al., 2023b)
Claude Opus 4.1	Anthropic	Transformer + Constitutional AI	Foco em segurança e alinhamento	(BAI et al., 2022)
DeepSeek V3.1 DeepThink	DeepSeek AI	Transformer	Alto desempenho em código e raciocínio	(DeepSeek-AI, 2024)
Gemini 2.5 Pro	Google	Transformer multimodal	Processamento multimodal	(Gemini Team, Google, 2023)

Fonte: Elaborado pelo autor.

A GenAI constitui uma classe de modelos computacionais cujo objetivo central é aprender a estrutura subjacente de um conjunto de dados e, a partir desse aprendizado, gerar novas amostras que preservem propriedades estatísticas relevantes dos dados originais. Diferentemente de abordagens discriminativas, que se concentram na previsão de rótulos ou na classificação de instâncias, os modelos generativos buscam estimar a distribuição de probabilidade dos dados, permitindo não apenas a análise, mas também a síntese de novas observações plausíveis (AKHTAR, 2024; GOODFELLOW et al., 2014).

2.1.1.1 Modelagem probabilística e geração de dados

Do ponto de vista formal, modelos generativos procuram aproximar a distribuição de probabilidade $p(x)$ dos dados observados, de modo que seja possível amostrar novas instâncias a partir dessa distribuição. Em contraste, modelos discriminativos estimam diretamente $p(y \mid x)$, focando na relação entre entrada e saída, sem necessariamente capturar a estrutura completa dos dados (GOODFELLOW et al., 2020).

Essa distinção é fundamental, pois a capacidade de geração de novas amostras de dados decorre diretamente da modelagem probabilística. Ao aprender padrões recorrentes em grandes volumes de dados, os modelos generativos tornam-se capazes de produzir conteúdos sintéticos, como, por exemplo, textos, imagens ou códigos. E, embora não correspondam a exemplos reais específicos, mantêm coerência estatística com o conjunto de treinamento (AKHTAR, 2024; BARRETO et al., 2023). Nesse sentido, a geração não implica criatividade no sentido humano, mas sim a recombinação de padrões aprendidos sob uma perspectiva probabilística.

2.1.1.2 Representações latentes e aprendizado de padrões

Um elemento central na modelagem generativa é o uso de representações latentes, que consistem em variáveis não observáveis utilizadas para capturar a estrutura subjacente dos dados. Em vez de operar diretamente no espaço original, frequentemente de alta dimensionalidade, os modelos aprendem projeções para um espaço latente mais compacto, no qual relações semânticas e estruturais podem ser representadas de forma mais eficiente (BENGIO; COURVILLE; VINCENT, 2013; KINGMA; WELLING, 2013; BOMMASANI et al., 2021).

Essas representações permitem ao modelo generalizar a partir de exemplos observados, identificando regularidades que transcendem instâncias específicas. No contexto de dados complexos, como linguagem natural, esse processo possibilita a captura de relações semânticas, sintáticas e contextuais, ainda que de forma implícita (ZHAO et al., 2023; NAVEED et al., 2023). Assim, o aprendizado em modelos generativos não se limita à memorização de exemplos, mas envolve a construção de abstrações estatísticas que permitem a geração de novas combinações coerentes.

2.1.1.3 Arquiteturas generativas

Diversas arquiteturas foram propostas para implementar modelos generativos, destacando-se, entre as mais relevantes, as *Generative Adversarial Networks* (GAN)s e os *Variational Autoencoder* (VAE)s. Essas abordagens diferem na forma como modelam a distribuição dos dados e no mecanismo utilizado para geração de novas amostras (GOODFELLOW et al., 2020; KINGMA; WELLING, 2013; AKHTAR, 2024).

As GANs operam por meio de um processo competitivo entre dois modelos (um gerador e um discriminador) no qual o gerador busca produzir amostras indistinguíveis dos dados reais, enquanto o discriminador tenta diferenciá-las (GOODFELLOW et al., 2014). Esse mecanismo *adversarial* conduz à melhoria progressiva da qualidade das amostras geradas, sendo particularmente eficaz na geração de dados de alta fidelidade.

Por sua vez, os VAEs adotam uma abordagem probabilística explícita baseada em inferência variacional, modelando a geração de dados a partir de variáveis latentes (KINGMA; WELLING, 2013). Nesse contexto, o modelo aprende simultaneamente uma função de

codificação, que projeta os dados no espaço latente, e uma função de decodificação, que reconstrói os dados a partir dessas representações. Essa estrutura permite não apenas a geração de novas amostras, mas também a interpretação do processo generativo em termos probabilísticos.

Embora essas arquiteturas tenham sido inicialmente desenvolvidas para dados estruturados ou visuais, seus princípios fundamentais de modelagem probabilística, aprendizado de representações e geração a partir de distribuições constituem a base conceitual para o desenvolvimento de modelos mais avançados, incluindo aqueles aplicados à linguagem natural (AKHTAR, 2024; NAVEED et al., 2023).

2.1.1.4 Limitações fundamentais dos modelos generativos

Apesar de seu potencial, modelos generativos apresentam limitações estruturais decorrentes de sua própria natureza probabilística. Em particular, esses modelos não possuem compreensão semântica ou causal do mundo, operando exclusivamente com base em correlações estatísticas aprendidas a partir dos dados de treinamento (BENDER et al., 2021).

Como resultado, a coerência das amostras geradas está associada à sua verossimilhança estatística, e não necessariamente à sua veracidade factual. Além disso, a qualidade das saídas está diretamente condicionada à distribuição dos dados utilizados no treinamento, podendo refletir ou amplificar vieses presentes nesses dados (BOMMASANI et al., 2021).

Outro aspecto relevante refere-se à ausência de mecanismos intrínsecos de validação factual. Como os modelos são otimizados para maximizar a probabilidade de sequências plausíveis, eles podem gerar conteúdos que, embora linguisticamente coerentes, sejam factualmente incorretos ou inconsistentes (JI et al., 2023). Essa limitação torna-se particularmente crítica em contextos que envolvem tomada de decisão, nos quais a distinção entre plausibilidade e veracidade é fundamental.

2.1.2 Grandes Modelos de Linguagem

Os LLMs representam uma evolução significativa dos modelos generativos aplicados ao processamento de linguagem natural, caracterizando-se pela capacidade de aprender padrões linguísticos complexos a partir de grandes volumes de dados textuais (ZHAO et al., 2023; NAVEED et al., 2023). Esses modelos são tipicamente baseados em arquiteturas profundas e treinados em larga escala, o que lhes permite capturar regularidades sintáticas, semânticas e contextuais da linguagem de forma abrangente.

Diferentemente de abordagens tradicionais em processamento de linguagem natural, que frequentemente dependem de regras explícitas ou representações estruturadas, os LLMs tratam a linguagem como uma sequência de unidades discretas (*tokens*) e modelam a probabilidade de ocorrência dessas sequências (BROWN et al., 2020). Nesse contexto,

a geração de texto consiste na predição iterativa do próximo *token* mais provável, dado um contexto previamente observado, refletindo a natureza probabilística desses modelos.

2.1.2.1 Representação da linguagem e *embeddings*

Para possibilitar o processamento computacional da linguagem, os LLMs utilizam representações vetoriais conhecidas como *embeddings*, nas quais cada *token* é mapeado para um vetor em um espaço contínuo de alta dimensionalidade. Essas representações permitem capturar relações semânticas e sintáticas entre palavras, de modo que *tokens* com significados semelhantes tendem a ocupar regiões próximas nesse espaço (NAVEED et al., 2023).

A construção desses espaços vetoriais é necessária para o funcionamento dos modelos, pois permite que operações matemáticas sejam utilizadas para inferir relações linguísticas e contextuais. Dessa forma, os *embeddings* constituem a base sobre a qual os modelos realizam tarefas de compreensão e geração de texto, ainda que essa "compreensão" seja, em essência, uma abstração estatística derivada dos dados de treinamento. Esses *embeddings* podem ser compreendidos como uma forma específica de representação latente, adaptada ao domínio da linguagem, na qual unidades discretas são projetadas em um espaço vetorial contínuo.

2.1.2.2 Arquitetura Transformer e mecanismo de atenção

A capacidade dos LLMs de capturar dependências contextuais de longo alcance decorre, em grande parte, da adoção da arquitetura *Transformer*, proposta por (VASWANI et al., 2017). Diferentemente de modelos anteriores baseados em recorrência *Recurrent Neural Networks* (RNN) ou convolução *Convolutional Neural Networks* (CNN), o *Transformer* utiliza mecanismos de atenção para modelar relações entre todos os elementos de uma sequência de forma simultânea.

O mecanismo de *self-attention* permite que cada *token* atribua pesos de relevância aos demais *tokens* da sequência, possibilitando a construção de representações contextuais dinâmicas. Esse processo é realizado por meio de múltiplas cabeças de consultas (*multi-head attention*) que permitem ao modelo capturar diferentes aspectos das relações linguísticas em paralelo. Como resultado, os LLMs conseguem integrar informações distribuídas ao longo do texto, melhorando sua capacidade de coerência e contextualização.

2.1.2.3 Processo de treinamento e modelos fundacionais

Os LLMs são tipicamente treinados em duas etapas principais: pré-treinamento e ajuste fino (*fine-tunning*). No pré-treinamento, o modelo é exposto a grandes corpora de texto, aprendendo padrões gerais da linguagem por meio de tarefas auto-supervisionadas,

como a predição do próximo *token* (BROWN et al., 2020). Essa etapa resulta em um modelo com conhecimento amplo, porém genérico.

Posteriormente, o modelo pode ser adaptado para tarefas específicas por meio de técnicas de ajuste fino, incluindo o uso de *feedback* humano, como no *Reinforcement Learning from Human Feedback* (RLHF) (OUYANG et al., 2022). Esse processo visa alinhar o comportamento do modelo com expectativas humanas, melhorando sua utilidade prática e reduzindo respostas indesejadas.

Essa abordagem é característica dos modelos fundacionais, que são treinados em larga escala e posteriormente adaptados para múltiplas aplicações (BOMMASANI et al., 2021). Esses modelos apresentam capacidades emergentes, como aprendizado em poucos exemplos (*few-shot learning*), permitindo sua aplicação em uma ampla variedade de tarefas sem necessidade de treinamento adicional específico.

2.1.2.4 Capacidades e aplicações dos LLMs

Devido à sua capacidade de generalização, os LLMs têm sido aplicados em diversas tarefas, incluindo geração de texto, tradução automática, resposta a perguntas e apoio à tomada de decisão em diferentes domínios (HADI et al., 2023). No contexto da segurança da informação, esses modelos têm sido explorados em atividades como identificação de vulnerabilidades, análise de incidentes, geração de relatórios técnicos e suporte a processos de avaliação de risco (FERRAG et al., 2024b; GENNARI et al., 2024).

Entretanto, embora essas aplicações demonstrem o potencial dos LLMs, seu desempenho depende fortemente da qualidade dos dados de treinamento e da adequação do contexto fornecido, o que pode impactar diretamente a confiabilidade das respostas geradas.

2.1.2.5 Limitações estruturais dos LLMs

Apesar de suas capacidades, os LLMs apresentam limitações inerentes à sua natureza estatística. Em particular, esses modelos não possuem compreensão semântica ou causal do mundo, operando com base em padrões probabilísticos aprendidos a partir de dados textuais. Como destacado por Bender et al. (2021), os modelos podem produzir respostas linguisticamente coerentes sem necessariamente refletir conhecimento factual ou entendimento real.

Um dos principais desafios associados aos LLMs é o fenômeno de alucinação, no qual o modelo gera informações incorretas ou não verificáveis com alto grau de confiança (JI et al., 2023). Esse comportamento decorre do fato de que o modelo é otimizado para maximizar a plausibilidade estatística das sequências geradas, e não sua veracidade.

Além disso, os LLMs podem reproduzir ou amplificar vieses presentes nos dados de treinamento, comprometendo a imparcialidade e a confiabilidade das respostas (BOMMASANI et al., 2021). A ausência de mecanismos explícitos de validação e a sensibilidade

a variações no contexto de entrada tornam esses modelos suscetíveis a inconsistências, especialmente em tarefas que exigem julgamento crítico.

Além dos fenômenos tradicionalmente descritos como alucinação, estudos recentes têm utilizado o termo "*delusional AI*" para descrever situações em que modelos geram respostas com alto grau de confiança, apesar de incorretas ou não verificáveis. Esse comportamento pode ser entendido como uma manifestação combinada de alucinação e excesso de confiança, na qual a plausibilidade linguística da resposta mascara a ausência de fundamento factual (JI et al., 2023; BENDER et al., 2021; BOMMASANI et al., 2021; CHANDRA et al., 2026; CHENG et al., 2026).

Outro aspecto importante refere-se à limitada interpretabilidade dos LLMs. Devido à complexidade de suas arquiteturas e ao elevado número de parâmetros, esses modelos são frequentemente considerados caixas-pretas, dificultando a compreensão dos fatores que influenciam suas respostas.

Nesse contexto, a *Explainable Artificial Intelligence* (xAI) busca desenvolver técnicas que permitam tornar os processos decisórios dos modelos mais transparentes e interpretáveis. No entanto, a aplicação dessas abordagens em LLMs ainda apresenta desafios significativos, especialmente em tarefas complexas que envolvem raciocínio contextual e geração de linguagem (MERSHA et al., 2024; BOMMASANI et al., 2021).

2.1.2.6 Implicações para tomada de decisão

As características estruturais dos LLMs têm implicações relevantes para seu uso em contextos que envolvem tomada de decisão. Embora esses modelos sejam capazes de gerar respostas plausíveis e coerentes, sua dependência de padrões estatísticos e ausência de entendimento causal limitam sua capacidade de fornecer avaliações fundamentadas em termos de risco, pois os LLMs operam predominantemente com base em associações estatísticas entre padrões linguísticos, não possuindo um modelo explícito de relações causais. Dessa forma, embora sejam capazes de produzir respostas plausíveis, essas respostas não necessariamente refletem uma compreensão dos mecanismos de causa e efeito subjacentes ao cenário de riscos analisado. Isso os diferencia fundamentalmente de uma análise de segurança da informação realizada por um especialista humano, capaz de raciocinar causalmente, entender mecanismos envolvidos de maneira ampla e ser capaz de avaliar as consequências.

Nesse sentido, a utilização de LLMs em atividades como avaliação de risco em segurança da informação deve ser analisada com cautela, considerando não apenas suas capacidades, mas também suas limitações. A distinção entre plausibilidade linguística e veracidade factual torna-se particularmente crítica, uma vez que decisões baseadas em informações imprecisas podem comprometer processos organizacionais relevantes, como implementar ou não um controle técnico que mitiga um determinado risco.

2.1.2.7 Vantagens e desafios éticos

Os LLMs surgiram como uma ferramenta poderosa em segurança da informação, influenciando significativamente as capacidades defensivas e ofensivas. No lado defensivo, eles auxiliam na análise de volumes massivos de registros de eventos de segurança e dados de tráfego de rede, identificando eficientemente anomalias, prevendo padrões de ataque e automatizando operações de segurança de rotina. Abordagens recentes demonstram que a combinação de LLMs com a análise tradicional de código estático melhora significativamente a detecção de vulnerabilidades, destacando o papel promissor das tecnologias de GenAI no aprimoramento da resiliência da segurança da informação (LI; DUTTA; NAIK, 2024).

Por outro lado, o potencial uso indevido de LLMs por cibercriminosos levantou alarmes significativos na comunidade de segurança, uma vez que adversários podem explorar a tecnologia para elaborar e-mails de *phishing* altamente persuasivos ou automatizar a descoberta e exploração de vulnerabilidades de software em grande escala (GENNARI et al., 2024). Além disto, existem preocupações éticas, incluindo o potencial de gerar desinformação, reforçar vieses existentes ou propagar involuntariamente estereótipos inerentes aos seus dados de treinamento. Esses modelos herdam vieses dos vastos volumes de dados nos quais são treinados, necessitando de rigorosos protocolos de avaliação e diretrizes éticas para mitigar potenciais danos e garantir a implementação responsável das tecnologias de GenAI (BENDER et al., 2021; AKHTAR, 2024; HADI et al., 2023).

Estas oportunidades e ameaças destacam a importância crítica de uma avaliação cuidadosa e de um planejamento estratégico ao implementar LLMs em áreas sensíveis como a segurança da informação. As pesquisas contínuas sobre as implicações práticas e éticas da GenAI são passos essenciais para uma integração tecnológica equilibrada e responsável (AKHTAR, 2024; GENNARI et al., 2024). Além disto, a escolha e o estabelecimento de *Frameworks* sólidos e abrangentes para a avaliação de riscos de segurança da informação com o uso de GenAI são importantes etapas no processo de gestão de riscos.

2.2 Gestão de Riscos de Segurança da Informação

A gestão de riscos em segurança da informação envolve a articulação entre estruturas formais, processos analíticos e julgamento humano. *Frameworks* de segurança fornecem diretrizes e modelos para a identificação e tratamento de riscos, enquanto as abordagens de análise de risco oferecem métodos para sua avaliação. No entanto, a aplicação desses instrumentos ocorre em contextos marcados por incerteza, nos quais o julgamento humano desempenha papel central na interpretação de cenários, na estimativa de probabilidades e na tomada de decisão.

2.2.1 *Frameworks* para Segurança da Informação

Os *Frameworks* são importantes na proteção de organizações contra ameaças cibernéticas cada vez mais sofisticadas. Essas ferramentas estruturadas fornecem uma base consistente para identificação, mitigação e gerenciamento de riscos em diversos setores. Além de oferecerem diretrizes claras, ajudam a alinhar as práticas organizacionais com as melhores práticas globais de segurança (MCINTOSH et al., 2024).

2.2.1.1 *CIS Framework*

O CIS é um conjunto de práticas e controles de segurança desenvolvido pelo *Center for Internet Security*, organização sem fins lucrativos voltada à promoção de boas práticas em cibersegurança. O objetivo do **Framework** é auxiliar organizações na proteção de seus ativos digitais por meio de recomendações práticas, priorizadas e operacionalmente orientadas, permitindo reduzir vulnerabilidades e mitigar ameaças cibernéticas (Center for Internet Security, 2021). O CIS foi originalmente criado a partir da consolidação de conhecimentos e experiências de especialistas da indústria, órgãos governamentais e profissionais da área de segurança da informação. Diferentemente de abordagens mais amplas e conceituais, como normas voltadas à governança e conformidade, o CIS possui forte foco operacional, descrevendo controles técnicos e administrativos que podem ser implementados diretamente pelas organizações. Essa característica o tornou um *Framework* amplamente adotado em ambientes corporativos, governamentais e acadêmicos como referência prática para fortalecimento da postura de segurança (GROŠ, 2019).

Na versão 8.1 (atual), o CIS é estruturado em 18 controles principais, abrangendo diferentes domínios da segurança da informação, como gerenciamento de ativos, controle de acesso, gerenciamento contínuo de vulnerabilidades, proteção contra *Malware*, monitoramento de atividades, resposta a incidentes e conscientização em segurança. Cada controle é subdividido em salvaguardas (*Safeguards*), que representam ações específicas destinadas à implementação prática das recomendações propostas pelo *Framework* (Center for Internet Security, 2021).

Outra característica importante do CIS é a utilização dos chamados *Implementation Groups (IGs)*, que permitem adaptar os controles de acordo com o porte, maturidade e perfil de risco das organizações. Os grupos são divididos em diferentes níveis de complexidade e maturidade, possibilitando que pequenas empresas, organizações de médio porte e ambientes críticos implementem controles compatíveis com suas necessidades e capacidades operacionais (Center for Internet Security, 2021).

Nesta dissertação, o CIS foi adotado como base para a construção dos cenários de avaliação de risco devido à sua natureza prática, prescritiva e amplamente utilizada na indústria de segurança da informação. Além disso, a estrutura padronizada dos 18 controles favorece a redução de ambiguidades na interpretação dos cenários, permitindo com-

parações mais consistentes entre avaliações realizadas por participantes humanos e por LLMs.

2.2.1.2 Outros *Frameworks*

Além do CIS, existem outros *frameworks* e normas amplamente utilizados para apoiar a gestão da segurança da informação e a análise de riscos em ambientes organizacionais. Entre os mais reconhecidos destacam-se o *NIST Cybersecurity Framework* e a norma ISO/IEC 27001.

O *NIST Cybersecurity Framework* (CSF), desenvolvido pelo *National Institute of Standards and Technology* (NIST) é um *Framework* reconhecido que oferece uma abordagem flexível e adaptável baseada em cinco funções principais: Identificar, Proteger, Detectar, Responder e Recuperar. Essas funções formam um ciclo de melhoria contínua, permitindo que as organizações abordem proativamente as ameaças cibernéticas e otimizem seus processos de segurança. Devido à sua ampla aplicabilidade e apoio governamental, o CSF é amplamente adotado por empresas privadas e instituições públicas (National Institute of Standards and Technology (NIST), 2024).

A Norma para Sistemas de Gestão de Segurança da Informação (ISO/IEC 27001), publicada pela Organização Internacional de Normalização, descreve os requisitos para a implementação de um Sistema de Gerenciamento de Segurança da Informação (SGSI). Essa norma é particularmente valorizada por seu foco na governança de segurança da informação, abrangendo desde políticas de controle de acesso até auditorias regulares. A certificação ISO/IEC 27001 tornou-se um padrão global, demonstrando o compromisso de uma organização com a proteção da informação e construindo confiança com clientes e parceiros ((ISO), 2022).

Para além de serem ferramentas operacionais, os *Frameworks* possuem um significado estratégico para a segurança da informação, pois servem como referência para conformidade regulatória e promovem a colaboração intersetorial, compartilhando melhores práticas e lições aprendidas. Ao adotar *Frameworks* para segurança da informação, as organizações podem aprimorar consistentemente a sua postura de segurança, fortalecendo assim a rede global de segurança da informação (FERRAG et al., 2024b).

2.2.2 Análise de Riscos

A análise de risco é um dos elementos na tomada de decisão em contextos caracterizados por incerteza, sendo amplamente utilizada em diversas áreas. De forma clássica, Kaplan e Garrick (1981) definem risco a partir de três componentes fundamentais: os cenários possíveis, a probabilidade de ocorrência de cada cenário e as respectivas consequências. Essa definição estabelece a base para abordagens quantitativas de risco, permitindo sua modelagem formal, ao mesmo tempo em que incorpora implicitamente a presença de

incerteza na identificação de cenários, na estimativa de probabilidades e na avaliação de impactos.

Abordagens mais recentes, como a de Aven e Renn (2009), ampliam essa conceituação ao tornar explícito o papel da incerteza, propondo que o risco seja compreendido como a combinação entre consequências e a incerteza associada aos eventos, destacando que a ausência de conhecimento completo limita a precisão das estimativas. Nesse contexto, a análise de risco não se restringe a cálculos probabilísticos, mas envolve também julgamentos subjetivos e interpretações baseadas em evidências disponíveis.

No domínio da segurança da informação, a avaliação de riscos apresenta desafios adicionais devido à natureza dinâmica das ameaças, à escassez de dados históricos confiáveis e à dependência de conhecimento especializado. *Frameworks* como a Norma para Gestão de Riscos (ISO/IEC 31000) ISO/IEC (2018) e o *Guide for Conducting Risk Assessments* (NIST/SP 800-30) NIST (2012) propõem processos estruturados para identificação, análise e tratamento de riscos, enquanto abordagens como o *Factor Analysis of Information Risk* (FAIR) buscam introduzir maior rigor quantitativo na estimativa de perdas. Apesar desses avanços, a análise de risco permanece, em grande medida, um processo dependente de julgamento humano, sujeito a vieses cognitivos e variações entre especialistas. Esse aspecto torna-se particularmente relevante ao considerar o uso de sistemas baseados em inteligência artificial como suporte à decisão, levantando questões sobre consistência, confiabilidade e alinhamento com a avaliação humana.

2.2.3 Julgamento humano e percepção de riscos

A avaliação de riscos envolve não apenas modelos formais, mas também processos cognitivos humanos que influenciam a interpretação de cenários incertos. Estudos clássicos sobre tomada de decisão demonstram que indivíduos frequentemente utilizam heurísticas ou atalhos mentais, para lidar com a complexidade e a incerteza, o que pode levar a vieses. Entre os principais vieses identificados estão a heurística da disponibilidade, na qual eventos mais facilmente lembrados são percebidos como mais prováveis, e a ancoragem, que leva indivíduos a se basearem excessivamente em informações iniciais (TVERSKY; KAHNEMAN, 1974). A Teoria da Perspectiva de Kahneman e Tversky (1979), evidencia que a avaliação humana de risco não segue estritamente princípios probabilísticos, sendo influenciada pela forma como ganhos e perdas são percebidos. Em particular, indivíduos tendem a superestimar eventos de baixa probabilidade e subestimar eventos mais prováveis, além de apresentar maior sensibilidade a perdas do que a ganhos equivalentes. Complementarmente, estudos sobre percepção de risco indicam que fatores emocionais e contextuais desempenham papel relevante na avaliação de ameaças. Slovic (1987) argumenta que o risco percebido é frequentemente moldado por características qualitativas, como familiaridade, controle percebido e potencial catastrófico, diferindo significativamente de avaliações baseadas em dados objetivos. Essa perspectiva é ampliada por Slovic

et al. (2004), que destacam a coexistência de dois modos de avaliação: um analítico e outro afetivo, sendo este último frequentemente dominante em situações de incerteza.

A experiência profissional também exerce influência significativa sobre a avaliação de risco. Especialistas tendem a desenvolver modelos mentais mais estruturados e maior capacidade de reconhecimento de padrões, o que pode resultar em julgamentos mais consistentes. No entanto, mesmo especialistas permanecem suscetíveis a vieses cognitivos, especialmente em contextos complexos e dinâmicos, como a segurança da informação. Nesse contexto, a análise de riscos em segurança da informação pode ser compreendida como um processo híbrido, que combina elementos técnicos e subjetivos. A variabilidade observada entre diferentes níveis de experiência, bem como a influência de vieses cognitivos, reforça a relevância de investigar o papel de sistemas automatizados, como modelos de linguagem de grande porte, no suporte à tomada de decisão em cenários de risco.

Sob essa óptica, embora a literatura clássica, notadamente a partir de Kahneman e Tversky, evidencie que o julgamento humano em contextos de incerteza é frequentemente influenciado por heurísticas e vieses cognitivos, essa abordagem, isoladamente, não é suficiente para explicar a variabilidade observada na avaliação de riscos em ambientes complexos. Complementarmente, a perspectiva da expertise, conforme proposta por Ericsson, Krampe e Tesch-Roemer (1993), demonstra que o desempenho superior decorre de longos períodos de prática deliberada, nos quais o indivíduo desenvolve estruturas cognitivas mais sofisticadas, capazes de reconhecer padrões, antecipar cenários e reduzir a dependência de processos heurísticos simplificadores. No contexto da segurança da informação, essa distinção torna-se particularmente relevante, uma vez que a identificação e a avaliação de ameaças dependem fortemente da capacidade do analista em interpretar sinais ambíguos, correlacionar eventos e tomar decisões sob pressão. Assim, enquanto indivíduos com menor experiência tendem a recorrer mais intensamente a heurísticas, e portanto tornando-se potencialmente mais suscetíveis a erros, profissionais com maior nível de expertise, construído ao longo de prática estruturada, apresentam maior acurácia e consistência na percepção e análise de riscos. Dessa forma, a percepção de risco em segurança da informação deve ser compreendida como um fenômeno híbrido, resultante tanto de características cognitivas inerentes ao processamento humano quanto do nível de especialização adquirido, reforçando a necessidade de abordagens que integrem vieses cognitivos e desenvolvimento de expertise na análise de risco.

2.3 Considerações finais

A fundamentação desenvolvida neste capítulo evidencia que a avaliação de riscos em segurança da informação não pode ser compreendida como um processo puramente técnico, sendo intrinsecamente dependente de fatores cognitivos, contextuais e experienciais. Embora *frameworks* e métodos estruturados forneçam diretrizes essenciais para a gestão

de riscos, sua aplicação prática está condicionada à interpretação humana, sujeita a heurísticas, vieses e limitações de conhecimento. Nesse sentido, a literatura demonstra que a variabilidade nas avaliações de risco decorre não apenas dessas limitações cognitivas, mas também do nível de expertise dos indivíduos, construído ao longo da experiência e da prática deliberada. Esse aspecto reforça o caráter não determinístico da análise de riscos, especialmente em domínios complexos como a segurança da informação, onde a incerteza e a ausência de dados completos são predominantes. Paralelamente, o avanço de modelos de linguagem de grande escala introduz uma nova dimensão a esse cenário, ao possibilitar a automação parcial de processos analíticos e a ampliação da capacidade de processamento de informações. No entanto, tais modelos também apresentam limitações relevantes, o que indica que sua aplicação não substitui o julgamento humano, mas o complementa. Dessa forma, a avaliação de riscos emerge como um processo híbrido, no qual a integração entre métodos estruturados, julgamento humano e tecnologias baseadas em Inteligência Artificial se mostra fundamental. Essa perspectiva sustenta a necessidade de investigar, de forma crítica, o papel dos LLMs nesse contexto, especialmente no que se refere à sua capacidade de apoiar, e não simplesmente substituir, a tomada de decisão em cenários de risco.

Capítulo 3

Trabalhos relacionados

A literatura situada na interseção entre segurança da informação e IA tem se expandido rapidamente, abrangendo desde *Benchmarks* voltados à avaliação de LLMs e modelos especializados para tarefas técnicas até *frameworks* e ferramentas aplicadas à análise de risco em contextos organizacionais. Esse cenário revela a coexistência de duas vertentes principais. De um lado, encontram-se estudos que exploram o potencial de LLMs e outras abordagens baseadas em IA para atividades como detecção de ameaças, avaliação de capacidades técnicas, geração de conteúdo e apoio a tarefas específicas de segurança. De outro, observa-se uma tradição mais consolidada de métodos e *frameworks* de avaliação de risco e maturidade, em geral orientados à governança, conformidade e apoio à decisão em ambientes reais.

Apesar dos avanços em ambas as frentes, ainda é pouco explorada a integração entre elas. Em particular, os estudos orientados por IA tendem a concentrar-se em conhecimento factual, raciocínio técnico, classificação ou detecção, enquanto os trabalhos de análise de risco permanecem majoritariamente dependentes de julgamento humano e de estruturas tradicionais de avaliação. Como resultado, permanece pouco explorada a comparação metodológica entre julgamentos humanos e inferências produzidas por LLMs em cenários estruturados de análise de risco em segurança da informação.

Nesse contexto, esta dissertação adota uma abordagem experimental baseada em cenários estruturados a partir dos Controles CIS, comparando avaliações numéricas de risco produzidas por profissionais humanos e por múltiplos LLMs. Mais especificamente, são analisadas a correlação entre os julgamentos, a ocorrência de padrões de subestimação ou divergência e a viabilidade de abordagens híbridas Humano-IA para apoio à análise de risco. Ao deslocar o foco de tarefas estritamente factuais ou diagnósticas para a atribuição de pontuações de risco em um contexto próximo à auditoria e à Governança, Riscos e Conformidade (GRC), esta dissertação busca oferecer uma contribuição original para a literatura.

Com base nesse posicionamento, este capítulo apresenta os principais trabalhos relacionados e organiza sua discussão em torno de três grupos: (i) *Benchmarks* para avaliação

de capacidades de LLMs em segurança da informação; (ii) modelos e sistemas especializados para tarefas específicas do domínio; e (iii) *frameworks*, métodos e ferramentas voltados à análise de risco e maturidade em contextos organizacionais. Essa estrutura permite situar de forma clara a contribuição desta dissertação e evidenciar a lacuna que ela procura preencher.

3.1 Benchmarks para avaliação de capacidades de LLMs

O autor Bhusal et al. (2024) propõem o *SECURE*, um *Benchmark* abrangente para avaliar LLMs em tarefas de segurança da informação, com ênfase particular em Sistemas de controle industrial (ICS). O *Benchmark* abrange seis conjuntos de tarefas envolvendo interpretação de vulnerabilidades, raciocínio de exploração, abuso de interpretador e geração de consultoria técnica, com base em fontes como *Common Weakness Enumeration* (CWE), *Common Vulnerabilities and Exposures* (CVE), *Adversarial Tactics, Techniques, and Common Knowledge* (MITRE ATT e CK) e *Cybersecurity and Infrastructure Security Agency* (CISA). A avaliação de sete LLMs – incluindo GPT-4, Gemini e Llama – revelou lacunas de desempenho substanciais, sublinhando a importância de *Benchmark* específicos de domínio e realistas.

Em Veuthey et al. (2025), os autores apresentam o MEQA, um *Framework* de meta-avaliação destinado a aferir a qualidade de *Benchmarks* de perguntas e respostas (Q&A) para LLMs. Embora o *Framework* seja geral, o artigo o demonstra por meio de sua aplicação a *Benchmarks* de segurança da informação. O *Framework* é organizado em oito critérios e 44 subcritérios, abrangendo robustez à memorização, robustez a *Prompts*, desenho da avaliação, desenho do avaliador, reprodutibilidade, comparabilidade, validade e confiabilidade. Ao aplicá-lo com três avaliadores humanos independentes e com o GPT-4o como avaliador auxiliar, o estudo evidencia forças e fragilidades metodológicas nos Benchmarks examinados. Em síntese, os resultados indicam bom desempenho relativo em reprodutibilidade e comparabilidade, mas apontam limitações recorrentes em robustez a *Prompts* e confiabilidade, além de elevada variabilidade entre subcritérios.

Tihanyi et al. (2024) apresentam o *CyberMetric*, um *Benchmark* concebido para avaliar o conhecimento factual e técnico de LLMs em segurança da informação por meio de questões de múltipla escolha distribuídas em nove domínios. O *Dataset* foi construído com apoio de GPT-3.5 e *Retrieval-Augmented Generation* (RAG), a partir de fontes amplas e reconhecidas, como padrões NIST, artigos científicos, *Requests For Comments* (RFC)s e livros públicos, resultando em quatro versões com diferentes tamanhos. As versões menores foram validadas por especialistas humanos, o que reforça a confiabilidade do instrumento. Como contribuição, o *CyberMetric* oferece uma base consistente para comparar modelos e humanos em condições *closed-book*. Contudo, seu foco permanece concentrado na mensuração de conhecimento técnico e factual, sem abordar de forma direta tarefas de

julgamento contextual, análise organizacional ou avaliação estruturada de risco, aspectos que são centrais no presente trabalho.

3.2 Modelos e sistemas especializados para tarefas de segurança

Agrawal et al. (2024) propõem o CyberGen, uma metodologia que combina LLMs e grafos de conhecimento para gerar automaticamente perguntas e respostas voltadas à educação em segurança da informação. A abordagem utiliza triplas extraídas do grafo de conhecimento *AISecKG* para enriquecer os prompts enviados ao ChatGPT, comparando estratégias de *zero-shot*, *few-shot* e *ontology-guided prompting*. Como resultado, o trabalho produz o conjunto de dados CyberQ, com aproximadamente 4.000 pares de perguntas e respostas, posteriormente avaliados por especialistas humanos. Com isso, oferece uma alternativa escalável para apoiar o desenvolvimento de materiais instrucionais e atividades de treinamento em segurança da informação.

Os autores Ferrag et al. (2024a) apresentam o *SecurityBERT*, um modelo leve baseado em Bidirectional Encoder Representations from Transformers (BERT) para detecção de ameaças cibernéticas em ambientes *Internet of Things* (IOT) e *Industrial Internet of Things* (IIoT). A proposta combina a arquitetura *Transformer* com a técnica *Privacy-Preserving Fixed-Length Encoding* (PPFLE), utilizada para representar o tráfego de rede de forma estruturada e compatível com o processamento do modelo. Avaliado no conjunto de dados *Edge-IIoTset*, o *SecurityBERT* alcançou 98,2% de acurácia na identificação de 14 tipos de ataques, com tempo de inferência inferior a 0,15s em CPU e tamanho de modelo de 16,7 MB, características que o tornam adequado para dispositivos com recursos limitados.

Levi et al. (2025), os autores apresentam o *CyberPal.AI*, uma família de LLMs especializada em segurança da informação, treinada com o conjunto de instruções *SecKnowledge*, construído a partir de conhecimento de domínio e de um processo multifásico orientado por especialistas. Como complemento, o trabalho introduz o *SecKnowledge-Eval*, uma suíte abrangente de avaliação composta por tarefas desenvolvidas especificamente para o domínio de segurança da informação e por *Benchmarks* públicos da área. Os resultados mostram melhorias significativas em relação aos modelos de base, com ganhos médios de até 24%, evidenciando o potencial de instruções orientadas por especialistas para o aprimoramento de modelos aplicados a tarefas complexas de segurança da informação.

3.3 *Frameworks* e métodos estruturados de análise de risco

Calculated... (2020) propõem o *Cybersecurity Evaluation Tool* (CET), um instrumento de avaliação de maturidade voltado a Pequenas e médias empresas (PME)s e baseado no *NIST Cybersecurity Framework*. O CET consiste em uma pesquisa *on-line* com 35 questões, respondidas por líderes de tecnologia para autoavaliar a maturidade organizacional nas cinco categorias do *NIST*: *identify*, *protect*, *detect*, *respond* e *recover*. Além da avaliação, a ferramenta oferece relatórios e recomendações práticas, permitindo que pequenas organizações interpretem e apliquem orientações de segurança da informação mesmo em contextos com recursos e experiência limitados.

Ekstedt et al. (2023) apresentam o *Yacraf* (*Yet Another Cybersecurity Risk Assessment Framework*), um *Framework* quantitativo de avaliação de risco cibernético que integra modelagem de ameaças, árvores de ataque, um metamodelo estruturado e uma lógica específica para cálculo de risco. A proposta busca considerar de forma mais explícita a arquitetura dos sistemas de tecnologia e seu contexto organizacional no processo de avaliação, oferecendo maior suporte analítico e visual à tomada de decisão. Como contribuição, o *Yacraf* unifica elementos de análise baseada em modelos e avaliação quantitativa de risco em uma única estrutura, sendo ilustrado por meio de um caso de uso e complementado por discussões sobre experiências práticas em organizações reais.

Haastrecht et al. (2021) propõem uma abordagem de avaliação de risco cibernético baseada em ameaças, voltada às necessidades de pequenas e médias empresas. O trabalho parte do reconhecimento de que as PMEs frequentemente dispõem de poucos recursos, conhecimento limitado e baixa motivação para investir em segurança, o que compromete sua capacidade de lidar com incidentes cibernéticos. Com base na *Self-Determination Theory* (*SDT*), a proposta busca não apenas produzir uma avaliação correta dos riscos, mas também incentivar a adoção de medidas de proteção, por meio de recomendações priorizadas e acionáveis. Como contribuição, o estudo descreve os requisitos de dados necessários para viabilizar automação e apresenta uma aplicação prática com diferentes interfaces de usuário, demonstrando como dados organizacionais podem ser convertidos em orientações concretas para apoio à decisão.

Finalmente, McIntosh et al. (2024) avaliam diferentes *frameworks* de governança, risco e conformidade, incluindo *Control Objectives for Information and Related Technology* (COBIT) e Norma para Gestão de Inteligência Artificial (ISO/IEC 42001), com o objetivo de identificar oportunidades, riscos e requisitos regulatórios associados ao uso de LLMs. O estudo discute como essas referências podem apoiar organizações na estruturação de práticas de governança para sistemas baseados em inteligência artificial, especialmente em cenários que demandam maior transparência, responsabilização e aderência regulatória. Como contribuição, o trabalho oferece uma análise comparativa orientada

à adoção organizacional de LLMs, destacando o papel dos *frameworks* de segurança e governança na mitigação de riscos emergentes.

3.4 Síntese comparativa dos trabalhos relacionados

Embora os estudos analisados abordem diferentes problemas na interseção entre segurança da informação e IA, sua comparação direta não é imediata, uma vez que variam em objetivos, escopo e métodos de avaliação. Nesse sentido, a Tabela 2 apresenta uma síntese comparativa dos principais trabalhos discutidos ao longo do Capítulo 3, considerando as diferentes dimensões. A coluna "Tipo de abordagem" descreve a natureza principal de cada proposta, classificando os trabalhos como *benchmarks*, métodos, ferramentas, *frameworks* ou modelos especializados. A coluna "Domínio" identifica o contexto ou área de aplicação do estudo, como segurança cibernética geral, ambientes industriais (ICS), IOT/IIoT, GRC ou PMEs. Já a coluna "Tarefa principal" resume o objetivo central de cada abordagem, incluindo atividades como avaliação de capacidades de LLMs, geração de perguntas e respostas, detecção de ameaças, avaliação quantitativa de riscos ou suporte à tomada de decisão.

As duas últimas colunas da Tabela 2 representam dimensões particularmente relevantes para esta dissertação. A coluna "Avaliação com humanos" indica se o trabalho incorpora especialistas humanos no processo de validação, avaliação ou comparação dos resultados produzidos pelos modelos. Por sua vez, a coluna "Análise de risco" identifica se o estudo aborda explicitamente tarefas relacionadas à avaliação ou gestão de riscos em segurança da informação. A análise conjunta dessas dimensões permite observar que grande parte dos trabalhos recentes concentra-se em avaliação de capacidades técnicas, geração de conteúdo, detecção de ameaças ou raciocínio factual, com menor ênfase em processos estruturados de julgamento de risco envolvendo participação humana.

Observa-se ainda que poucos trabalhos combinam simultaneamente análise de risco, participação de especialistas humanos e comparação direta com Grandes Modelos de Linguagem (LLMs). Alguns estudos apresentam mecanismos quantitativos de avaliação de risco, enquanto outros utilizam validação humana ou exploram aplicações específicas de IA em segurança da informação. Entretanto, permanece limitada a quantidade de pesquisas que investigam como os LLMs se comportam em cenários de análise de risco organizacional que envolvem subjetividade, experiência profissional e percepção contextual de ameaças.

Tabela 2 – Comparação dos trabalhos relacionados

Trabalho	**Tipo de abordagem**	**Domínio**	**Tarefa principal**	**Avaliação com humanos**	**Análise de risco**
SECURE (Bhusal et al., 2024)	*Benchmark* de LLM	ICS / segurança	Avaliação multitarefa de raciocínio	Não	Não
MEQA (Veuthey et al., 2025)	Meta-avaliação	*Benchmarks* de LLM	Avaliação de métricas e critérios de qualidade	Sim	Não
CyberMetric (Tihanyi et al., 2024)	*Benchmark*	Conhecimento técnico em cibersegurança	Q&A de múltipla escolha	Não	Não
CyberGen (Agrawal et al., 2024)	Método / *dataset*	Educação em segurança	Geração de Q&A	Sim	Não
SecurityBERT (Ferrag et al., 2024)	Modelo especializado	IOT / IIoT	Detecção de ameaças	Não	Não
CyberPal.AI (Levi et al., 2025)	Família de LLMs especializada	Cibersegurança	Assistência e raciocínio em tarefas de segurança	Não	Parcial
CET (Benz & Chatterjee, 2020)	Ferramenta	PMEs / *NIST*	Avaliação de maturidade e recomendações	Sim	Sim
McIntosh et al. (2024)	*Framework* / avaliação comparativa	Governança e conformidade de LLMs	Avaliação de *frameworks* regulatórios e de governança	Não	Parcial
Yacraf (Ekstedt et al., 2023)	*Framework* quantitativo	Avaliação de risco cibernético	Avaliação quantitativa de risco	Não	Sim
GEIGER (van Haastrecht et al., 2021)	Método / ferramenta	PMEs	Pontuação de risco e recomendações	Sim	Sim
Este trabalho	Estudo experimental	GRC / *CIS Controls*	Avaliação de risco (1–10)	Sim	Sim

3.5 Considerações finais

A análise dos trabalhos relacionados apresentada neste capítulo evidencia que a aplicação de IA e LLM em segurança da informação tem recebido atenção crescente na literatura recente. Os estudos analisados demonstram avanços importantes em diferentes frentes,

incluindo a criação de *benchmarks* para avaliação de capacidades técnicas de LLMs, o desenvolvimento de modelos especializados para tarefas de segurança e a proposição de ferramentas e *frameworks* voltados à avaliação de riscos, maturidade e governança.

Apesar dessas contribuições, observa-se que muitos trabalhos permanecem concentrados em tarefas técnicas ou factuais, como resposta a perguntas, classificação de conhecimento, detecção de ameaças, análise de vulnerabilidades ou geração de conteúdo especializado. Outros estudos abordam a análise de risco de forma mais estruturada, porém geralmente sem investigar diretamente o comportamento de LLMs em comparação com o julgamento de profissionais humanos. Assim, até onde foi possível verificar na literatura analisada, ainda são limitados os estudos que comparam, de forma metodológica, avaliações humanas e avaliações produzidas por modelos de linguagem em cenários estruturados de análise de risco em segurança da informação.

Dessa forma, os trabalhos discutidos neste capítulo ajudam a delimitar a lacuna explorada por esta dissertação. Ao combinar cenários estruturados baseados nos controles CIS, participação de especialistas humanos e avaliação comparativa com múltiplos LLMs, este trabalho se posiciona na interseção entre segurança da informação, análise de riscos e IA.

Capítulo 4

Materiais e Métodos

Este capítulo apresenta o método utilizado para investigar a eficácia dos LLMs na avaliação de riscos de segurança da informação e comparar seu desempenho com o de profissionais humanos em diferentes níveis de experiência. Para atingir esse objetivo, foi concebida uma metodologia estruturada em quatro etapas principais, conforme detalhado na Seção 4.1. Estas fases seguem uma abordagem sequencial para garantir uma análise estruturada e rigorosa do processo de avaliação de risco. Na Seção 4.2, é detalhado o panorama das quatro etapas do método e depois uma especificação aprofundada de cada uma das etapas do método.

4.1 Panorama do método em quatro etapas

A) **Etapa I - Criação de Dados:** Esta fase envolve o desenvolvimento de dados de entrada contextualizados para orientar o processo de avaliação de risco. Definimos um contexto geral, especificando um cenário de segurança da informação no qual um analista de segurança da informação deve avaliar o risco de segurança de uma empresa parceira. Com base nisso, criamos um conjunto de perguntas estruturadas e respostas correspondentes para simular avaliações do mundo real (Cenários).

   - ❑ **Cenários** Nesta subetapa, foi desenvolvido um conjunto de perguntas baseadas no *Framework* CIS, de forma que cada uma das perguntas represente os 18 controles existentes. Com base na experiência prévia do pesquisador, foram criadas e testadas situações reais de segurança cibernética para elevar a precisão e a qualidade das respostas. A partir destas descobertas, foram elaboradas respostas alinhadas com um cenário simulado que representa uma empresa de tecnologia de médio porte.

B) **Etapa II - Testes com LLMs:** Nesta etapa, os dados da **Etapa I** (Criação de Dados) foram testados usando os modelos de linguagem selecionados. Alimentamos os modelos com respostas estruturadas para investigar a sua capacidade de avaliar

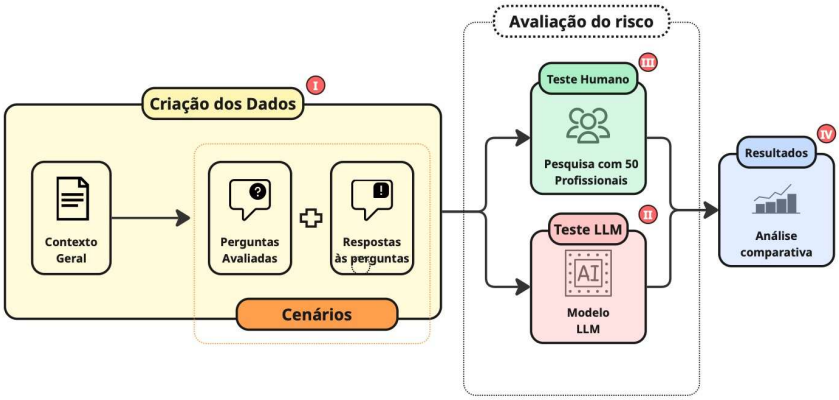

Figura 1 – Método para avaliação de riscos usando LLMs e avaliação humana.

riscos de segurança da informação. Diferentes variações da subetapa cenários foram testadas, incluindo casos extremos e mistos, para avaliar a consistência dos modelos na avaliação de riscos.

C) **Etapa III - Pesquisa com Humanos:** Foram convidados 50 participantes humanos com experiências diversas em Segurança da Informação e Tecnologia. Estes participantes avaliaram os riscos de segurança usando o mesmo cenário contextual utilizado na avaliação com LLMs. Preparamos um questionário estruturado para coletar suas avaliações e informações demográficas e de experiência profissional. Todos os participantes foram informados, por meio do formulário online utilizado para a coleta de dados, de que suas respostas anonimizadas seriam utilizadas exclusivamente para fins de pesquisa acadêmica. Ao preencherem e enviarem voluntariamente o formulário, os participantes forneceram seu consentimento livre e esclarecido para o uso de seus dados anonimizados neste estudo. De acordo com a regulamentação brasileira para pesquisas envolvendo seres humanos — especificamente a Resolução CNS nº 510/2016, que disciplina a supervisão ética de estudos nas áreas de ciências humanas e sociais —, atividades de pesquisa que coletam dados anônimos de opinião, sem qualquer possibilidade de identificação dos participantes, estão dispensadas de apreciação por Comitê de Ética em Pesquisa[1]. Portanto, não foi necessária aprovação de comitê de ética para este estudo.

D) **Etapa IV - Análise de Resultados:** Por fim, foi realizada a análise estatística das respostas tanto dos LLMs quanto dos participantes humanos. O estudo examinou a média das classificações de risco, a distribuição, a correlação de *Pearson* entre as avaliações humanas e de IA, e aplicou testes de significância estatística (Teste pareado) para identificar potenciais tendências ou discrepâncias na percepção de risco.

[1] Registro 98453826.7.0000.5152 no Comitê de Ética

4.2 Especificação das quatro etapas do método

4.2.1 Etapa I - Criação de Dados

Ao trabalhar com LLMs para geração de texto, é fundamental desenvolver um contexto geral que estabeleça o ambiente atual e descreva o que precisa ser feito, expondo o máximo de detalhes possível (ZHAO et al., 2023). Este contexto geral permite que o LLM compreenda o seu propósito real, reduzindo as chances de má interpretação. Além disso, o contexto é essencial para auxiliar os indivíduos na compreensão de um assunto específico (NAVEED et al., 2023). Assim, o contexto geral desenvolvido é:

"Você é um analista de segurança da informação trabalhando para uma empresa de médio porte. Esta empresa lida com informações confidenciais de clientes, desde dados de identificação pessoal até registros estratégicos. Atualmente, a empresa está preocupada com a postura de segurança da informação adotada por seus parceiros.

Estes parceiros desempenham um papel essencial na cadeia de suprimentos, fornecendo serviços que suportam processos e funções críticas de negócios. Eles exigem acesso a recursos de rede, incluindo aplicativos internos e externos, dados e serviços de tecnologia. Para avaliar os riscos associados a estes parceiros, a empresa utiliza o Framework CIS, que possibilita a identificação de riscos potenciais antes do início da prestação de serviços.

Atualmente, um dos parceiros está sendo avaliado quanto aos riscos de segurança da informação associados à sua infraestrutura, processos, sistemas, aplicações e outros aspectos. Os dados necessários para esta avaliação são apresentados abaixo, incluindo perguntas personalizadas alinhadas com os padrões CIS e as respostas do parceiro.

Seu objetivo, como analista de segurança da informação, é determinar o risco de segurança associado a cada parceiro com base na sua experiência, atribuindo uma classificação numa escala de 1 (baixo risco) a 10 (alto risco), para definir o nível de risco que cada parceiro representa para a empresa."

Com o objetivo de promover a transparência, a reprodutibilidade e a reutilização dos artefatos produzidos nesta pesquisa, os conjuntos de dados, cenários de avaliação e resultados consolidados utilizados neste estudo foram disponibilizados publicamente em um repositório *GitHub*[2].

4.2.1.1 Cenários

O processo de criação de cenários baseou-se nos detalhes descritos no contexto geral. Consistiu em dois elementos principais: as perguntas feitas pelo analista de segurança e as respostas fornecidas pelo parceiro avaliado. As entradas foram cuidadosamente concebidas para simular um cenário realista de avaliação de risco de segurança da informação, garantindo consistência e autenticidade no processo de avaliação.

[2] Ver <https://github.com/gusberto/cybersecurity_assessment_llm_hitl/blob/main/>

A versão 8 do *Framework* CIS foi a referência para o desenvolvimento das perguntas. Selecionamos um tópico de segurança da informação para cada um dos 18 controles que melhor representasse as práticas essenciais associadas a esse controle. Com base nesses tópicos, formulamos perguntas claras e objetivas que refletem o tipo de avaliação que um analista de segurança poderia usar no parceiro avaliado. Estas perguntas foram concebidas para capturar de forma abrangente os aspectos chave de cada controle, garantindo que a informação recebida fosse relevante e valiosa para a avaliação de risco.

As respostas simuladas dos parceiros foram elaboradas com o objetivo de representar, de forma verossímil, padrões frequentemente observados em avaliações práticas de segurança da informação conduzidas em contextos organizacionais. Essa construção foi orientada pela experiência profissional prévia do pesquisador em processos de análise e avaliação de segurança, sem utilização de dados identificáveis, documentos confidenciais ou informações específicas de organizações reais. Observou-se, a partir dessa experiência prática, que respostas fornecidas por parceiros e fornecedores tendem a apresentar uma visão predominantemente positiva sobre suas práticas de segurança, ao mesmo tempo em que reconhecem limitações operacionais, lacunas de cobertura ou desafios de implementação. Além disso, tais respostas costumam conter informações complementares além daquilo que foi explicitamente solicitado, o que contribui para reduzir ambiguidades interpretativas e aproximar os cenários experimentais de situações encontradas em avaliações reais de risco.

O conjunto final de entradas foi estruturado para incluir tanto as perguntas formuladas quanto as respostas simuladas. As respostas seguiram um padrão realista, predominantemente positivo, mas incorporando aspectos negativos relevantes para garantir o equilíbrio. Este padrão foi definido com base na análise de avaliações do mundo real, refletindo a forma como os parceiros de negócios normalmente respondem a tais avaliações. Esta abordagem permite-nos avaliar se o modelo de linguagem consegue identificar e ponderar os riscos com precisão, tal como faria um analista humano. O código-fonte está disponível no repositório do *GitHub*: <https://github.com/gusberto/cybersecurity_assessment_llm_hitl/blob/main/results_final_scenarios.xlsx> ou, a seguir:

1. **Tema:** Inventário de Ativos de Hardware
   Pergunta: A empresa possui um inventário detalhado de ativos corporativos?
   Resposta: A empresa mantém um inventário de ativos abrangente e detalhado. Este inventário cobre a grande maioria dos ativos da organização, com aproximadamente 98% contabilizados e rastreados. O inventário é continuamente atualizado através de sistemas automatizados, garantindo que novos ativos sejam capturados e os existentes sejam atualizados em tempo real à medida que ocorrem mudanças.

2. **Tema:** Inventário e Controle de Ativos de Software
   Pergunta: A empresa possui um inventário detalhado de ativos, sistemas e aplica-

ções de software?
Resposta: A organização possui um inventário de software bem estabelecido e mantido de forma eficaz. Atualmente, aproximadamente 85% dos softwares da organização estão inventariados, garantindo um controle abrangente sobre os ativos de software. Este inventário é atualizado regularmente através de processos manuais, garantindo que novas instalações e atualizações sejam refletidas no sistema.

3. **Tema:** Gerenciamento da Informação
Pergunta: Existe um processo formal de gerenciamento de dados implementado?
Resposta: Um processo formal de gerenciamento de dados está em vigor dentro da organização. Os dados são classificados por sensibilidade. O processo de gerenciamento de dados é auditado a cada seis meses, e as permissões de acesso são revisadas regularmente.

4. **Tema:** Processo de Configuração Segura (*hardening, security baselines*, etc.)
Pergunta: Existe um processo de configuração segura para todos os sistemas?
Resposta: Um processo de configuração segura está em vigor para todos os sistemas, baseado na estrutura CIS (Center for Internet Security). As configurações são padronizadas em toda a organização e são revisadas e atualizadas trimestralmente.

5. **Tema:** Inventário de Contas de Usuário
Pergunta: Existe um inventário de todas as contas de usuário em todos os sistemas listados no catálogo de aplicações?
Resposta: Um inventário de todas as contas de usuário do sistema no catálogo de aplicações é mantido. Este inventário é atualizado frequente e automaticamente. As contas são gerenciadas centralmente e as contas não utilizadas são prontamente desativadas.

6. **Tema:** Provisionamento de Acesso em Sistemas e Aplicações
Pergunta: Existe um processo formal para concessão de acesso?
Resposta: Um processo formal de concessão de acesso está em vigor. Neste processo, as solicitações são aprovadas com base no princípio do menor privilégio, e o acesso é revisado periodicamente para garantir que ainda é necessário. Além disso, os acessos concedidos são baseados em função ou cargo.

7. **Tema:** Gerenciamento de Vulnerabilidades
Pergunta: Existe um processo de gerenciamento de vulnerabilidades em vigor para todos os ativos (*endpoints*, servidores, dispositivos de rede, aplicações, sistemas, etc.)?
Resposta: Existe um processo de gerenciamento de vulnerabilidades para todos os ativos, incluindo *endpoints*, servidores, dispositivos de rede, aplicações e sistemas.

As vulnerabilidades são identificadas e priorizadas para remediação. As varreduras de vulnerabilidades são realizadas mensalmente e os relatórios de gestão são revisados regularmente.

8. **Tema:** Gerenciamento de *Log* e Registro de Atividades
   Pergunta: Existe um processo formal de gerenciamento de *logs* de auditoria implementado?
   Resposta: Um processo formal de gerenciamento de *logs* de auditoria está em vigor, com coleta automática de *logs* de todos os sistemas e armazenamento seguro em repositórios centralizados e criptografados. Equipes dedicadas revisam regularmente os logs usando ferramentas automatizadas para detectar atividades suspeitas. De acordo com a política, os *logs* de auditoria são retidos por 12 meses em um ambiente seguro, com acesso restrito a usuários autorizados.

9. **Tema:** Uso de Navegadores e Clientes de e-mail
   Pergunta: Todos os navegadores e clientes de e-mail são totalmente suportados, e apenas versões autorizadas são implementadas?
   Resposta: Todos os navegadores e clientes de e-mail em uso são totalmente suportados, com atualizações regulares e suporte contínuo do fornecedor. As configurações de segurança para navegadores e clientes de e-mail são aplicadas e monitoradas centralmente para garantir a conformidade. A lista de versões suportadas é revisada e atualizada trimestralmente para garantir que as versões mais recentes e seguras estejam em uso.

10. **Tema:** *Antimalware*
    Pergunta: O software *Antimalware* está implantado em todos os *endpoints*?
    Resposta: O software *Antimalware* está implantado em todos os *endpoints* organizacionais, com varreduras regulares de *malware* realizadas. A eficácia do software *antimalware* é testada trimestralmente. O recurso anti-*exploit* está ativado e os *endpoints* são gerenciados centralmente através de uma plataforma de segurança dedicada.

11. **Tema:** Recuperação de Dados e Informações
    Pergunta: Existe um processo de recuperação de dados implementado?
    Resposta: Um processo de recuperação de dados está em vigor, com backups realizados regularmente e armazenados de forma segura. Dados críticos podem ser restaurados dentro de 8 horas em caso de perda. Os logs de recuperação são mantidos e revisados periodicamente para garantir a integridade do processo.

12. **Tema:** Manter a Infraestrutura de Rede Atualizada
    Pergunta: A infraestrutura de rede é atualizada regularmente?
    Resposta: A infraestrutura de rede é atualizada regularmente, com *firmware* e

patches de software aplicados prontamente. Todas as alterações de configuração em dispositivos de rede são rastreadas e documentadas. A infraestrutura de rede é auditada quanto à conformidade semestralmente.

13. **Tema:** Centralização de Eventos de Segurança da Informação
    Pergunta: Todos os eventos de segurança da informação e relacionados são centralizados?
    Resposta: Todos os alertas de eventos de segurança são centralizados numa plataforma dedicada, permitindo o monitoramento em tempo real de toda a infraestrutura. Todos os eventos críticos são enviados para este sistema central, onde são analisados e tratados por equipes especializadas.

14. **Tema:** Programa de Conscientização
    Pergunta: Existe atualmente um programa para aumentar a conscientização ou treinar funcionários sobre tópicos de cibersegurança?
    Resposta: Um programa ativo de conscientização de segurança está em vigor, e todos os funcionários devem completar o treinamento de conscientização de segurança. O treinamento é realizado anualmente.

15. **Tema:** Gerenciamento de Terceiros, Parceiros e Fornecedores
    Pergunta: Como é realizado atualmente o gerenciamento de terceiros, parceiros e fornecedores em relação ao mapeamento e análise de riscos de segurança da informação?
    Resposta: Um inventário de todos os prestadores de serviços é mantido e atualizado regularmente. Os prestadores de serviços são revisados periodicamente quanto à conformidade de segurança e suas práticas de segurança são avaliadas anualmente.

16. **Tema:** Desenvolvimento Seguro de Software
    Pergunta: Práticas seguras de desenvolvimento de software são implementadas?
    Resposta: Um Processo de Desenvolvimento Seguro de Aplicações está estabelecido, com práticas de segurança incorporadas em todas as fases do ciclo de vida de desenvolvimento. As equipes de desenvolvimento seguem padrões de codificação segura e conduzem revisões de segurança regulares, como testes de vulnerabilidade e auditorias de código. Atualizações e *patches* de segurança são aplicados continuamente para garantir a proteção das aplicações.

17. **Tema:** Resposta a Incidentes
    Pergunta: Existe um processo de resposta a incidentes em vigor?
    Resposta: Um Processo de Resposta a Incidentes está em vigor, com uma equipe designada responsável pelo tratamento de incidentes. O treinamento de tratamento de incidentes é realizado regularmente e todos os incidentes são documentados e revisados.

18. **Tema:** Testes de Penetração e Simulações de Ataque
    Pergunta: Testes de penetração, testes de segurança ou simulações de ataque são realizados?
    Resposta: Um Programa de Testes de Penetração está em vigor, com testes regulares realizados em sistemas críticos e infraestrutura de rede. Equipes especializadas realizam estes testes para identificar e explorar vulnerabilidades, garantindo que as ações corretivas sejam tomadas prontamente.

4.2.2 Testes com LLMs

Antes de aplicar o modelo de linguagem ao conjunto de dados principal desenvolvido para este trabalho, foi conduzida uma análise preliminar para verificar se os LLMs selecionados poderiam gerar respostas coerentes e contextualmente fundamentadas no contexto de avaliações de risco. Para isso, foram utilizadas as mesmas perguntas do conjunto de dados principal, mas com um novo conjunto experimental de respostas simplificadas, conforme ilustrado na Figura 2.

Foram projetados quatro cenários distintos, sendo que cada pergunta recebeu quatro variações de respostas:

1. Cenário 1: Todas as respostas como “Sim” para todas as perguntas.
2. Cenário 2: Todas as respostas como “Não” para todas as perguntas.
3. Cenário 3: Respostas mistas, mas predominantemente “Sim”.
4. Cenário 4: Respostas mistas, mas predominantemente “Não”.

Desta forma, usando os testes preliminares foi possível avaliar como os LLMs interpretam cenários extremos e moderados, e verificar se conseguiam reconhecer padrões e fornecer avaliações justificadas, em vez de respostas superficiais ou arbitrárias. Os resultados[3] demonstraram que todos os modelos interpretaram com precisão as variações de resposta e atribuíram classificações de risco coerentes, sem recorrer a avaliações extremas ou inconsistentes. Consequentemente, isto confirmou que todos os cinco modelos (ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro) eram adequados para aplicação neste estudo.

Após esta fase preliminar, iniciou-se a avaliação principal utilizando o conjunto de dados[4] do cenário 3, explicitamente concebido para permitir comparações entre os LLMs e os especialistas humanos em segurança da informação. O modelo analisou cada resposta

[3] Ver *GitHub*:<https://github.com/gusberto/cybersecurity_assessment_llm_hitl/blob/main/llm_tested_scenarios.xlsx>

[4] Ver *GitHub*:<https://github.com/gusberto/cybersecurity_assessment_llm_hitl/blob/main/results_final_scenarios.xlsx>

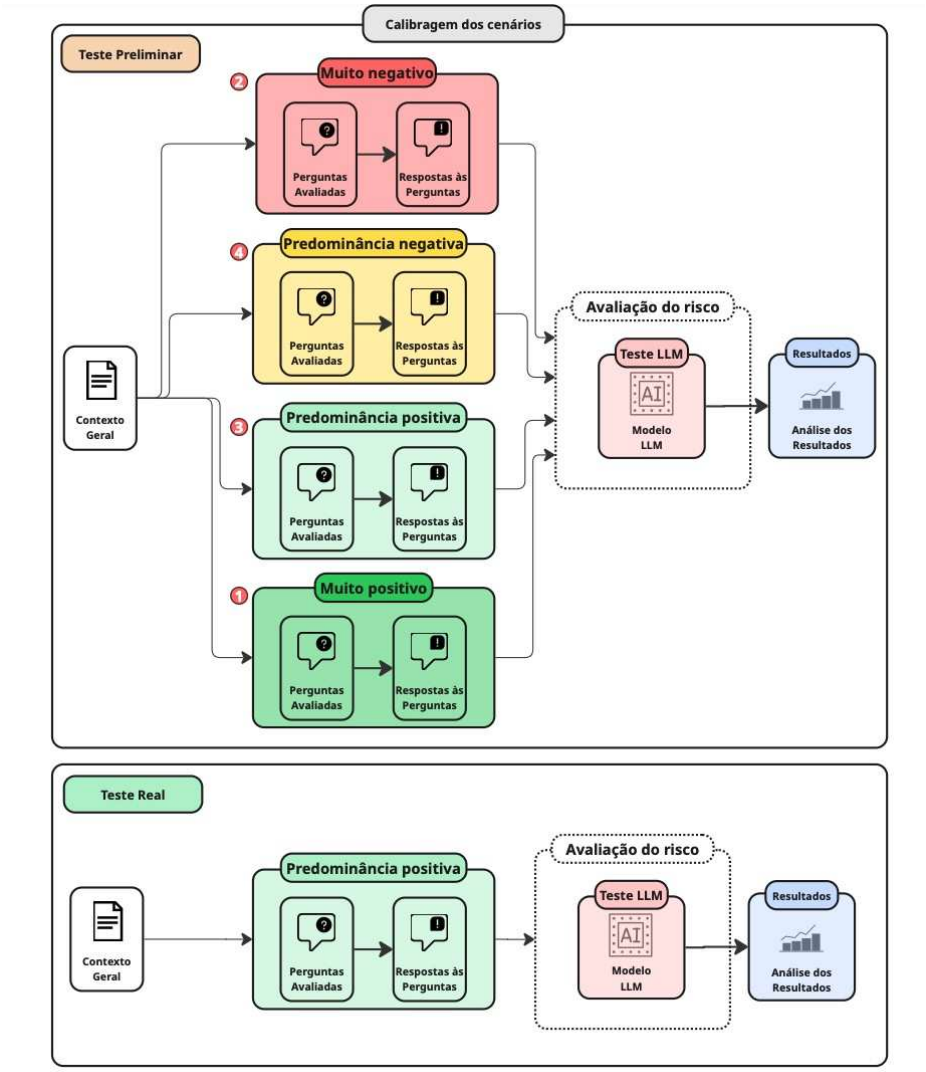

Figura 2 – Fluxograma ilustrando a diferença entre o teste preliminar e o teste real.

fornecida, atribuindo um nível de risco e justificativas detalhadas para cada avaliação. Além disso, foi solicitado aos modelos que fornecessem uma classificação de risco geral para o cenário completo. Para garantir a consistência, repetimos os testes em diferentes conversas e contas, verificando a estabilidade dos resultados. Em todas as execuções, os LLMs demonstraram consistência nas suas classificações e justificativas, destacando a sua capacidade analítica no contexto da avaliação de risco de segurança da informação.

Para minimizar potenciais variações nas respostas e garantir maior robustez estatística, cada avaliação foi realizada cinco vezes em conversas independentes para cada modelo. Consequentemente, os valores utilizados na análise final são a média dos cinco testes realizados por modelo, proporcionando uma visão mais equilibrada e confiável do desempenho dos LLMs.

4.2.3 Etapa III - Pesquisa com Humanos

Foram convidados 50 participantes através das redes sociais e acadêmicas, bem como instituições de pesquisa e empresas de tecnologia para se voluntariarem neste estudo. Semelhante ao teste com LLMs, esta fase de Teste Humano envolveu o recrutamento de pessoas para preencherem um formulário no mesmo contexto fornecido ao LLM e classificarem os cenários de acordo com o nível de risco que representavam para a empresa. Para isto, desenvolvemos um formulário do Google, disponível no Apêndice A, que, além dos cenários, continha perguntas demográficas, tais como a área de especialização dos participantes, tipo de perfil, anos de experiência e nível profissional. Por fim, os participantes classificaram o nível de risco dos cenários e justificaram, opcionalmente, o risco atribuído, conforme artefato[5] disponibilizado.

Os participantes variaram de indivíduos que tinham entrado recentemente na área (seja em segurança ou em tecnologia em geral) a profissionais com mais de 10 anos de experiência comprovada e certificações internacionalmente reconhecidas. Além disso, para garantir respostas imparciais na análise, foi solicitado aos voluntários que preenchessem o questionário sem assistência externa. Após o encerramento da coleta, todas as respostas foram revisadas manualmente para verificar sua completude, coerência e aderência à escala de avaliação proposta. Não foram identificadas respostas inconsistentes ou que justificassem a exclusão de participantes da amostra.

4.2.4 Etapa IV - Análise dos resultados

As respostas da pesquisa e as respostas dos LLMs foram analisadas usando estatísticas descritivas para identificar padrões de resposta, com foco na média, dispersão e distribuição das classificações de risco atribuídas pelos participantes e pelos cinco LLMs distintos. Para aprofundar a compreensão das diferenças entre os grupos, segmentamos os participantes com base no nível profissional e organizamos a análise em três agrupamentos:

- ❏ Grupo 0: Todos os Participantes
- ❏ Grupo 1: Profissionais altamente especializados (Sênior e Especialistas)
- ❏ Grupo 2: Profissionais menos especializados (Nível Inicial e Intermediário)

Em paralelo, comparamos as classificações atribuídas pelos LLMs com as médias dos participantes humanos para identificar discrepâncias e possíveis tendências nas avaliações dos modelos. Além disso, comparamos os LLMs para determinar qual teve melhor desempenho no contexto da avaliação de risco de segurança da informação. Esta comparação fornece descobertas sobre os pontos fortes e fracos relativos de cada LLM, destacando

[5] Ver *GitHub*:<https://github.com/gusberto/cybersecurity_assessment_llm_hitl/blob/main/pesquisa-de-riscos-ciberneticos-respostas.csv>

quais abordagens podem ser mais eficazes para aplicações práticas de segurança da informação.

Para verificar a significância das diferenças observadas, aplicamos os testes estatísticos descritos no Capítulo 5 para determinar se as diferenças entre os grupos eram estatisticamente significativas. Além disso, analisamos a correlação linear entre as respostas dos participantes e as avaliações dos modelos para determinar se ambos seguiam um padrão semelhante na percepção de risco. Também examinamos a distribuição das pontuações atribuídas para compreender o intervalo e a frequência dos valores fornecidos por humanos e pelos LLMs. Esta abordagem ajuda a identificar diferenças e a determinar se elas ocorrem de forma sistemática ou aleatória.

Para garantir a validade dos resultados, adotamos diversas medidas para minimizar vieses e ampliar a diversidade da amostra. O recrutamento dos participantes incluiu profissionais de diferentes níveis e áreas dentro da Segurança da Informação, fornecendo uma gama diversificada de perspectivas. Além disso, o questionário foi administrado anonimamente e sem influência externa, reduzindo vieses nos julgamentos individuais. Por fim, a robustez da análise foi reforçada por técnicas estatísticas rigorosas, que permitiram uma interpretação confiável dos resultados e contribuíram para uma melhor compreensão das diferenças na percepção de risco entre humanos e IA.

4.3 Considerações finais

Este capítulo apresentou o método experimental desenvolvido para investigar o uso de Grandes Modelos de Linguagem na avaliação de riscos de segurança da informação. Diferentemente de abordagens focadas em tarefas de classificação, detecção ou geração de conteúdo, o método proposto foi concebido para avaliar a capacidade dos modelos em reproduzir julgamentos de risco realizados por profissionais humanos em cenários estruturados de segurança da informação. Como parte desse processo, foi construído um conjunto de 18 cenários fundamentados nos controles do CIS, permitindo a criação de um *benchmark* padronizado para comparação entre avaliações humanas e avaliações produzidas por LLMs. Além disso, foi estabelecido um procedimento preliminar de calibração para verificar a consistência dos modelos diante de diferentes variações dos cenários. A metodologia também descreveu o processo de recrutamento dos participantes humanos, a caracterização da amostra, a forma de coleta das respostas e o processo de obtenção das avaliações dos cinco modelos investigados. O principal resultado deste capítulo é a construção de uma base metodológica e experimental que possibilita analisar, de forma objetiva, o alinhamento entre avaliações de risco realizadas por especialistas humanos e por Grandes Modelos de Linguagem. A partir desse conjunto de dados e do método experimental desenvolvido, o próximo capítulo apresenta os resultados obtidos e discute em que medida os modelos investigados se aproximam ou divergem do julgamento humano

na avaliação de riscos de segurança da informação.

Capítulo 5

Experimentos e Análise dos Resultados

No Capítulo 5 são apresentados os resultados do experimento conduzido ao longo desta dissertação, juntamente com as respostas às perguntas de pesquisa propostas. Os resultados estão organizados da seguinte forma: na Seção 5.1, descrevemos as características do experimento, enquanto na Seção 5.2 abordamos a pesquisa com os participantes humanos, incluindo informações como o seu nível profissional, anos de experiência, área de atividade principal e perfil profissional. Então, nas Seções 5.3 e 5.4 respondemos às perguntas de pesquisa com os dados e achados do experimento, respectivamente. Por fim, apresentamos a discussão e as limitações da pesquisa na Seção 5.5.

5.1 Visão Geral dos Experimentos

Nesta pesquisa, os participantes apresentados na Seção 5.2 e os cinco modelos apresentados na Tabela 1 da Seção 2.1.1 do Capítulo 2, atribuíram uma classificação de risco em uma escala de 1 a 10 para os 18 cenários base, pertinentes a cada um dos controles do CIS, apresentados no Apêndice A, usando uma escala contínua numérica. Para fins analíticos, a escala de risco de 1 a 10 foi agrupada em três faixas interpretativas: baixo risco, para pontuações de 1 a 3; risco médio, para pontuações de 4 a 6; e alto risco, para pontuações de 7 a 10. Ao comparar as pontuações e examinar como as classificações desses cinco modelos se correlacionam com as dos participantes humanos, obtemos evidências sobre a tendência geral de percepção de risco entre humanos e os modelos LLMs. O coeficiente de correlação de *Pearson* foi utilizado para mensurar a direção e a intensidade da associação linear entre as avaliações atribuídas pelos LLMs e pelos participantes humanos, e adotou-se a escala proposta por Schober, Boer e Schwarte (2018), utilizada como critério operacional para classificar a força da associação linear observada. Segundo essa classificação, valores absolutos de correlação entre 0,00 e 0,09 indicam associação desprezível; entre 0,10 e 0,39, associação fraca; entre 0,40 e 0,69, associação moderada; entre 0,70 e 0,89, associação forte; e entre 0,90 e 1,00, associação muito forte. Essa escala foi adotada neste estudo por oferecer uma referência objetiva para a interpretação dos

coeficientes, permitindo maior consistência na comparação entre os diferentes modelos e grupos de participantes. Ainda assim, a interpretação dos resultados considera o contexto experimental da pesquisa, especialmente o fato de que correlação positiva indica alinhamento direcional entre as avaliações, mas não implica equivalência entre os julgamentos de risco. Ao longo do capítulo, são destacadas as diferenças estatísticas (como a tendência dos cinco modelos de pontuar mais baixo do que os participantes humanos) e examina-se se as classificações de risco relativas dos modelos correspondem aos padrões observados entre os respondentes humanos.

O objetivo não foi alcançar consenso absoluto entre os participantes humanos, mas capturar a distribuição natural das percepções de risco entre profissionais com diferentes níveis de experiência e especialização. Essa variabilidade reflete a realidade organizacional da avaliação de riscos em segurança da informação e constitui um dado relevante para a análise comparativa com os modelos de linguagem.

5.2 Dados Demográficos dos Participantes Humanos

No período de 28 de novembro de 2024 a 9 de dezembro de 2024, foram convidados 50 participantes para esta pesquisa através de redes sociais e acadêmicas, grupos de discussão privados e contato direto com indivíduos de universidades, instituições de pesquisa e empresas. A pesquisa foi criada usando um formulário do Google sem a identificação dos participantes e incluiu perguntas sobre os anos de experiência dos participantes, nível profissional, perfil profissional e área de especialização. Foi solicitado aos voluntários — variando de iniciantes a especialistas — que preenchessem o questionário sem assistência externa.

Para melhor compreender as características dos participantes da pesquisa, foi realizada uma análise estatística dos seus atributos demográficos e profissionais. O conjunto de dados inclui informações sobre anos de experiência, nível profissional, perfil de trabalho e área de especialização. Esta análise fornece informações sobre a distribuição do conhecimento entre os respondentes, destacando variações nos níveis de experiência e representação da indústria.

As Tabelas 3, 4, 5 e 6 resumem a distribuição dos participantes nas principais categorias demográficas e profissionais. Estas visões oferecem uma compreensão mais clara da composição da amostra.

Os participantes foram agrupados em quatro categorias, cada uma definida por atributos específicos destinados a melhor compreender sua percepção de risco nos 18 cenários da pesquisa:

1. Anos de experiência: Anos de experiência destes profissionais na área de conhecimento de Tecnologia ou Segurança da Informação;

Tabela 3 – Anos de experiência

Anos de experiência	**Total**	**%**
10 ou mais anos	25	50%
4 anos	8	16%
5 ou mais anos	6	12%
3 anos	4	8%
2 anos	4	8%
Até 1 ano	2	4%
1 ano	1	2%

Tabela 4 – Área de especialidade

Área de especialidade	**Total**	**%**
Tecnologia (Infraestrutura)	11	22%
Segurança da Informação (Defensiva)	11	22%
Tecnologia (Software)	9	18%
Segurança da Informação (Gerenciamento de Riscos)	5	10%
Tecnologia (Dados)	4	8%
Outras	3	6%
Segurança da Informação (Governança)	3	6%
Tecnologia (Governança)	2	4%
Segurança da Informação (Ofensiva)	2	4%

Tabela 5 – Perfil profissional

Perfil profissional	**Total**	**%**
Técnico	22	44%
Ambos (Técnico e Gerencial)	17	34%
Gerencial	11	22%

Tabela 6 – Nível Profissional

Nível Profissional	**Total**	**%**
Sênior	19	38%
Especialista	13	26%
Intermediário	10	20%
Outros	4	8%
Estagiário	3	6%
Júnior	1	2%

2. Área de especialidade: Em qual sub-área de conhecimento de Tecnologia ou Segurança da Informação eles estão trabalhando atualmente;

3. Perfil profissional: Profissionais mais focados no âmbito técnico, gerencial ou a combinação de ambos;

4. Nível profissional: A posição que ocupam, com base nas habilidades, conhecimentos ou experiência que possuem atualmente, o que lhes confere autoridade e capacidade de tomada de decisão, descritas na forma dos atuais cargos.

A pesquisa, que se concentra na atribuição de níveis de risco a situações de segurança da informação, baseia-se na experiência de 50 profissionais, dos quais 50% têm mais de 10 anos de experiência ou ocupam cargos de Sênior ou Especialista. Além disso, a experiência diversa dos participantes — abrangendo Segurança da Informação, Software, Dados, Infraestrutura e Governança — e os seus papéis variados, sejam técnicos, gerenciais ou a mistura de ambos, contribuem para uma compreensão abrangente de como diferentes experiências contribuem para a avaliação de riscos.

5.3 PP 1: LLMs vs. Humanos (todos)

Em que medida se pode confiar em LLMs para conduzir avaliações de risco de segurança da informação sem o envolvimento de especialistas humanos?

A pergunta de pesquisa 1 visa avaliar se os grandes modelos de linguagem, como os cinco modelos investigados nessa dissertação e detalhados na Tabela 1 (ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro), podem ser utilizados com confiança para avaliar riscos de segurança da informação sem o envolvimento humano. Para isso, comparamos o seu desempenho com todos os participantes humanos pesquisados num conjunto de cenários padronizados.

5.3.1 Visão geral dos riscos

Primeiramente, analisamos os dados comparando o risco atribuído pelos cinco modelos com aquele atribuído por todos os participantes humanos. Para cada pergunta do cenário, calculamos o risco atribuído pelos modelos, o risco médio atribuído pelos participantes humanos e o desvio padrão das respostas. A Tabela 7 mostra o risco médio atribuído a cada questão pelos cinco modelos (ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro) versus todos os participantes humanos pesquisados.

Tabela 7 – Visão geral dos riscos | LLMs vs. Humanos

P	**ChatGPT**	**DeepSeek**	**Llama**	**Claude**	**Gemini**	$\overline{Humanos}$	**DP**
P1	1.60	0.60	1.80	1.00	0.90	3.70	2.45
P2	5.30	3.40	3.60	3.20	5.80	5.50	2.34
P3	2.60	1.60	1.80	2.00	1.60	3.40	2.40
P4	2.20	0.60	1.60	1.00	1.00	3.70	2.66
P5	2.00	0.60	1.40	1.00	0.50	2.90	2.33
P6	2.90	1.20	1.60	1.60	0.80	3.10	2.25
P7	4.60	2.40	1.80	2.00	2.10	4.50	2.74
P8	2.30	0.60	1.20	1.00	1.00	3.00	2.12
P9	2.30	1.00	1.80	1.20	1.00	3.60	2.59
P10	2.30	0.80	1.40	1.00	1.00	3.50	2.70
P11	3.60	1.80	2.20	2.60	2.20	3.80	2.44
P12	2.60	1.60	1.60	1.80	1.50	3.40	2.40
P13	2.20	0.80	1.20	1.00	1.00	3.00	2.06
P14	4.00	3.80	2.40	3.80	3.60	3.80	2.61
P15	3.80	2.60	3.40	2.80	3.10	4.00	2.30
P16	2.70	1.20	1.60	1.20	1.20	3.30	2.53
P17	2.80	1.60	1.60	1.60	1.20	3.30	2.44
P18	2.80	0.80	1.60	1.40	1.00	3.30	2.48
Total	**2.92**	**1.50**	**1.87**	**1.73**	**1.69**	**3.60**	**2.44**

Para esta primeira etapa, a análise dos dados mostra que o risco médio geral (média aritmética das respostas) atribuído pelos participantes humanos é maior do que o atribuído pelos cinco modelos para todas as questões. O risco médio (média aritmética das respostas) atribuído pelos modelos para as 18 questões é de 1.94, enquanto o risco médio atribuído pelos participantes humanos para as mesmas 18 questões é de 3.60. Os participantes humanos atribuem um risco mais elevado com um desvio padrão de 2.44. Na média geral entre os cinco modelos e os participantes humanos, os modelos estimam um risco 1.66 mais baixo do que os participantes humanos. A Figura 3 demonstra o gráfico de dispersão em que é possível observar como se comportam os LLMs e humanos em relação aos riscos médios atribuídos para cada cenário.

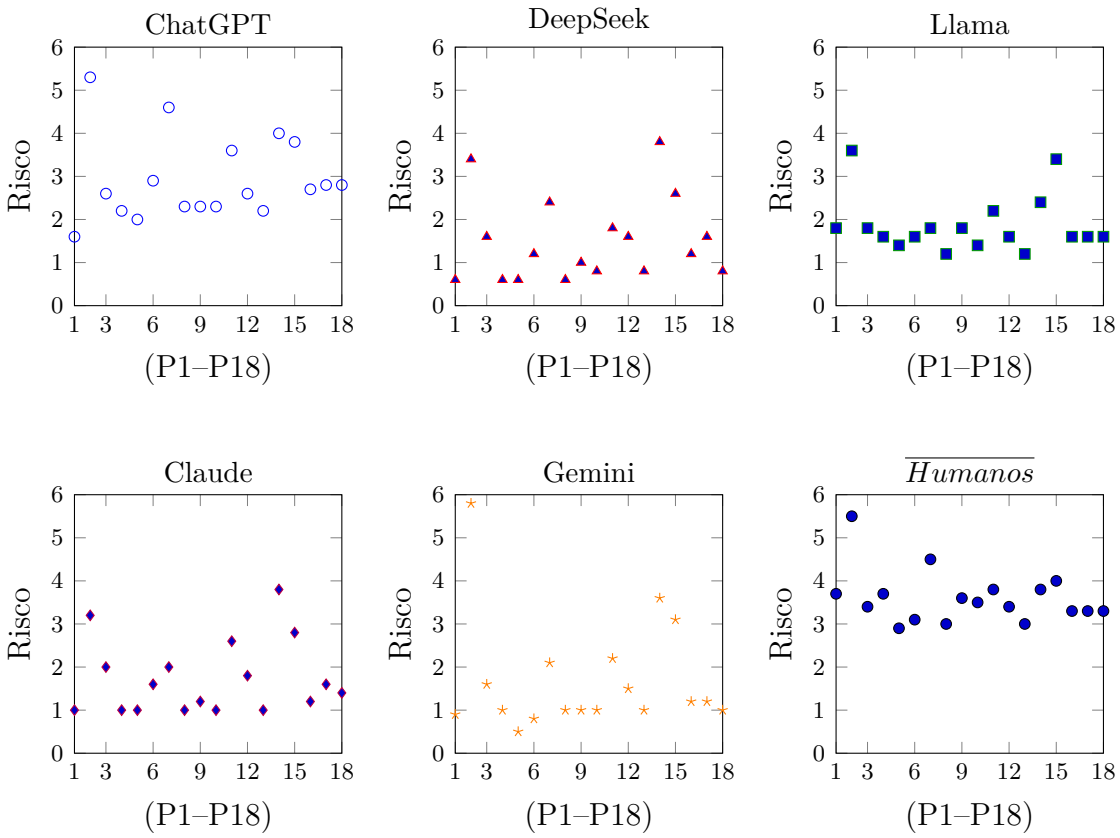

Figura 3 – Dispersão de Risco x Perguntas | LLMs vs. Humanos

5.3.2 Quantificação da associação LLMs vs. Humanos

Segundo, quantificamos a relação entre os cinco modelos e os participantes humanos. Em termos práticos, existe uma associação direta entre as classificações de risco atribuídas pelos modelos e as dos participantes humanos? Ou seja, as classificações de risco movem-se na mesma direção? A fim de examinar essa relação, utilizamos o coeficiente de correlação de *Pearson*[1] para medir a força da relação linear entre duas variáveis. Se houver uma forte relação linear, o coeficiente de correlação está próximo de 1 ou -1; 0 indica que não há relação linear.

O coeficiente de correlação de *Pearson* para cada um dos cinco modelos, demonstrado na Tabela 8, indica associação positiva entre as avaliações produzidas pelos LLMs e aquelas atribuídas pelos participantes humanos. Com base na classificação de Schober, Boer e Schwarte (2018), foram observadas correlações positivas fortes para ChatGPT (r = 0,817), DeepSeek (r = 0,706), Llama (r = 0,801) e Gemini (r = 0,858), enquanto Claude apresentou correlação positiva moderada (r = 0,630). Esses resultados indicam que, embora os modelos atribuam pontuações absolutas de risco mais baixas, eles tendem a acompanhar a direção geral das avaliações humanas: quando os participantes humanos atribuem maior risco a determinado cenário, os modelos também tendem a elevar suas

[1] Ver Schober, Boer e Schwarte (2018), para discussão sobre uso e interpretação de coeficientes de correlação.

pontuações, ainda que em menor magnitude.

Tabela 8 – Coeficiente *Pearson* | LLMs vs. Humanos

	ChatGPT	**DeepSeek**	**Llama**	**Claude**	**Gemini**
$\overline{Humanos}$	0.817	0.706	0.801	0.630	0.858

5.3.3 Teste t

Terceiro, conduzimos testes estatísticos para avaliar a significância das diferenças usando um teste t[2] de amostras emparelhadas para os cinco modelos e participantes humanos. O teste t de amostras emparelhadas, também conhecido como teste t de amostras dependentes, é um teste estatístico para determinar se a diferença média entre dois conjuntos de dados de observação é igual a zero. Cada objeto ou entidade é medido duas vezes neste teste, resultando em dois conjuntos de observações. O teste t de amostras emparelhadas emprega duas hipóteses de pesquisa contraditórias, a hipótese nula e a hipótese alternativa, como na maioria dos procedimentos estatísticos. A hipótese nula afirma que não há diferença nas médias dos dois conjuntos de dados emparelhados. De acordo com esta perspectiva, todas as distinções visíveis resultam de variação aleatória. É possível que a diferença média real entre as duas amostras não seja igual a zero, ao contrário da hipótese alternativa.

Tabela 9 – Teste t emparelhado | LLMs vs. Humanos

Modelo	**Valor t**	**Valor p (bicaudal)**
ChatGPT	4.91	1.32×10^{-4}
DeepSeek	12.71	4.16×10^{-10}
Llama	17.59	2.41×10^{-12}
Claude	11.69	1.49×10^{-9}
Gemini	9.59	2.83×10^{-8}

Dado que todos os valores p obtidos nos testes t emparelhados são extremamente pequenos **(todos p<0.001)**, podemos rejeitar com segurança a hipótese nula em todos os casos. Este resultado indica que, apesar das correlações positivas entre as avaliações dos cinco modelos e as classificações humanas, existem diferenças estatisticamente significativas nas classificações de risco atribuídas pelos modelos em comparação com as dos participantes humanos. Em outras palavras, embora os modelos possam seguir tendências semelhantes, as suas avaliações de risco absolutas divergem estatisticamente daquelas fornecidas pelos humanos.

[2] Ver *Paired Sample t-test*, Talikan et al. (2025)

5.3.4 Distribuição da frequência

Por fim, os histogramas mostram a distribuição das classificações de risco atribuídas pelos cinco modelos versus as dadas pelos participantes humanos, conforme apresentado na Figura 6. Embora os participantes humanos tenham classificado predominantemente os cenários com risco entre 3 e 5, os LLMs concentraram as suas avaliações na faixa entre 1 e 3, com apenas valores mais altos ocasionais. Isso indica que, embora os LLMs sigam tendências semelhantes às dos humanos, eles subestimam o nível geral de risco, reforçando os resultados do teste t.

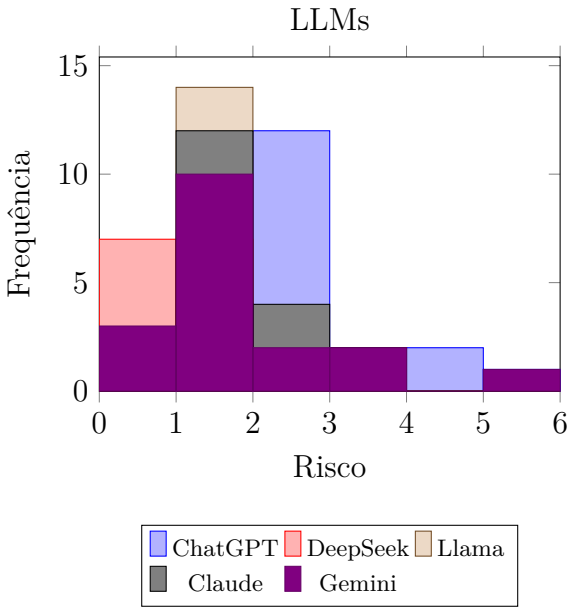

Figura 4 – Histograma de riscos | LLMs

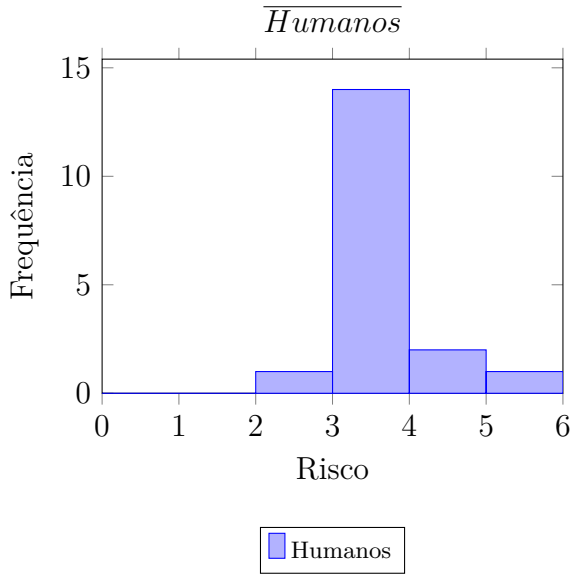

Figura 5 – Histograma de riscos | Humanos

Figura 6 – Histograma de riscos atribuídos | LLMs vs. Humanos

5.3.5 Considerações finais da pergunta de pesquisa 1

As descobertas discutidas na Seção 5.3 sustentam a conclusão de que os cinco modelos subestimam estatisticamente o risco de segurança da informação em comparação com a percepção de todos os participantes humanos. Além disso, o desvio padrão de 2.44 entre os participantes humanos indica uma distribuição significativa de opiniões. Embora alguns participantes possam classificar um cenário como risco muito baixo, outros podem classificá-lo como risco mais elevado. Esta variação sugere que o grupo de participantes humanos não tem um consenso rígido sobre os cenários, refletindo uma ampla gama de perspectivas ou interpretações ao julgar o risco, o que será melhor detalhado nas seções seguintes.

5.4 PP 2: LLMs vs. Humanos (por nível profissional)

Como os LLMs utilizados nesta dissertação se comparam aos profissionais humanos na avaliação de risco em cenários padronizados, analisados conforme o nível profissional?

Esta pergunta de pesquisa visa investigar em que medida os grandes modelos de linguagem podem replicar ou aproximar os julgamentos de profissionais de segurança da informação e tecnologia ao realizar avaliações de risco nos 18 cenários de risco. Um dos atributos utilizados nesta análise para agrupar os participantes humanos em categorias é o Nível Profissional, categoria que compreende dois grandes grupos de pessoas, sendo que os participantes seniores e especialista representam 64% da amostra e todos os outros níveis representam os 36% restantes. Dada a diversidade de expertise dentro da força de trabalho de tecnologia e segurança da informação, esta questão é subdividida em duas perguntas que refletem diferentes níveis de experiência profissional, conforme as Subseções 5.4.1 e 5.4.2:

5.4.1 PP 2.1: LLMs vs. Humanos (Seniores/Especialistas)

Como o desempenho dos cinco modelos se compara ao de profissionais seniores e especialistas, dos quais se espera um conhecimento técnico mais profundo e uma percepção de risco mais refinada?

5.4.1.1 Visão geral dos riscos

Primeiro, analisamos os dados comparando o risco atribuído pelos cinco modelos com aquele atribuído pelos participantes humanos, conforme demonstrado na Tabela 10. Desta vez, apenas o grupo de participantes com níveis profissionais Sênior e Especialista. Para cada questão do cenário, calculamos o risco atribuído por cada um dos cinco modelos, o risco médio atribuído pelos Profissionais Seniores e Especialistas, e o desvio padrão das suas respostas.

Para esta primeira etapa, a análise de dados mostra que o risco médio geral atribuído por este grupo de participantes humanos Seniores e Especialistas é mais alto do que o atribuído pelos cinco modelos em todas as questões, e mais alto do que o atribuído por todos os participantes. Além da média de risco de 3.87, observa-se um desvio padrão menor de 2.41 em comparação com o grupo que contém todos os participantes, indicando um melhor consenso dentro deste grupo. A Figura 7 demonstra o gráfico de dispersão que revela como se comportam os LLMs e os humanos Seniores e Especialistas em relação aos riscos médios atribuídos para cada cenário.

Tabela 10 – Visão geral dos riscos | LLMs vs. Humanos Seniores e Especialistas

P	**ChatGPT**	**DeepSeek**	**Llama**	**Claude**	**Gemini**	$\overline{Humanos}$	**DP**
P1	1.60	0.60	1.80	1.00	0.90	3.66	1.94
P2	5.30	3.40	3.60	3.20	5.80	6.22	2.05
P3	2.60	1.60	1.80	2.00	1.60	3.63	2.44
P4	2.20	0.60	1.60	1.00	1.00	3.97	2.53
P5	2.00	0.60	1.40	1.00	0.50	2.84	2.08
P6	2.90	1.20	1.60	1.60	0.80	3.47	2.35
P7	4.60	2.40	1.80	2.00	2.10	5.16	2.74
P8	2.30	0.60	1.20	1.00	1.00	3.25	2.32
P9	2.30	1.00	1.80	1.20	1.00	4.00	2.61
P10	2.30	0.80	1.40	1.00	1.00	3.88	2.87
P11	3.60	1.80	2.20	2.60	2.20	4.09	2.34
P12	2.60	1.60	1.60	1.80	1.50	3.53	2.35
P13	2.20	0.80	1.20	1.00	1.00	3.13	2.04
P14	4.00	3.80	2.40	3.80	3.60	4.16	2.80
P15	3.80	2.60	3.40	2.80	3.10	4.38	2.23
P16	2.70	1.20	1.60	1.20	1.20	3.28	2.46
P17	2.80	1.60	1.60	1.60	1.20	3.50	2.48
P18	2.80	0.80	1.60	1.40	1.00	3.59	2.61
Total	**2.92**	**1.50**	**1.87**	**1.73**	**1.69**	**3.87**	**2.41**

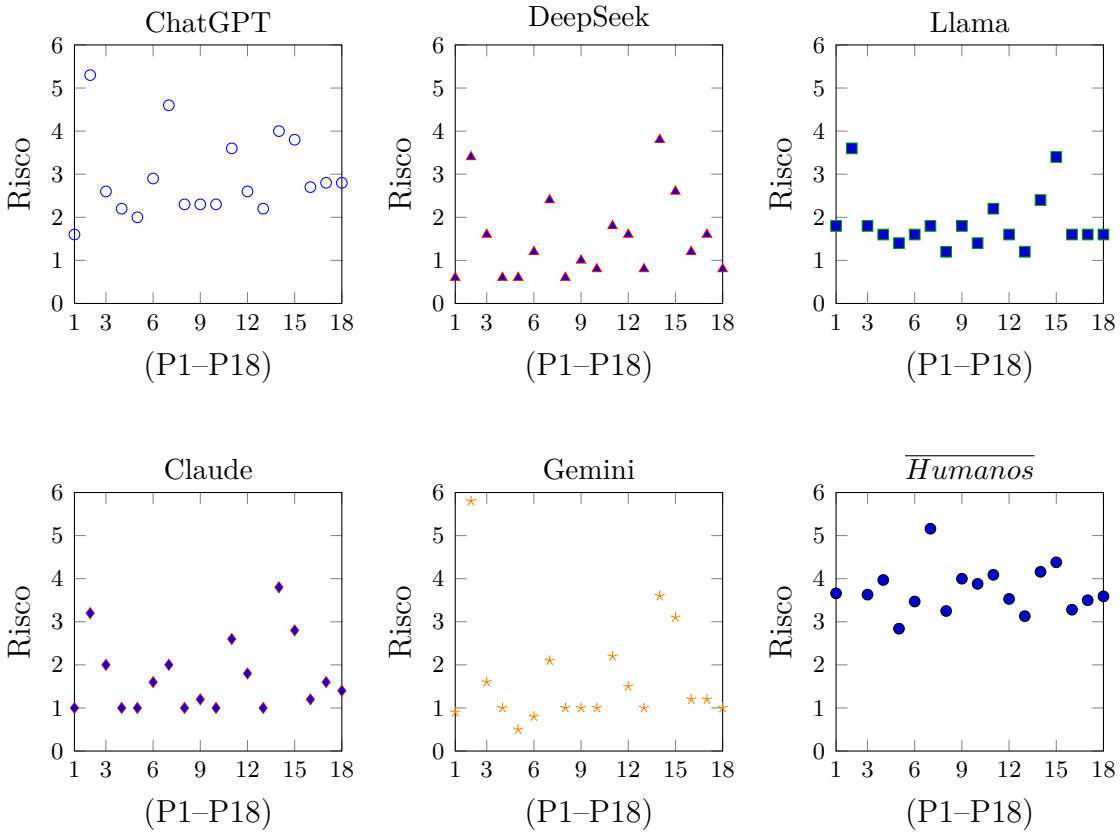

Figura 7 – Dispersão de Risco x Perguntas | LLMs vs. Humanos Seniores e Especialistas

5.4.1.2 Quantificação da associação LLMs vs. Humanos Seniores e Especialistas

Segundo, quantificamos a relação entre os cinco modelos (ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro) e as fornecidas pelos participantes humanos. Especificamente, procuramos avaliar se existe uma relação direta, ou seja, se as classificações de risco tendem a mover-se na mesma direção. Para examinar isso, empregamos o coeficiente de correlação de *Pearson*[3] para medir a força da relação linear entre os dois conjuntos de classificações. Um coeficiente próximo de 1 ou -1 indica uma forte relação linear, enquanto um valor próximo de 0 sugere nenhuma correlação linear.

O coeficiente de correlação de *Pearson* para cada um dos cinco modelos, demonstrado na Tabela 11, indica associação positiva entre as avaliações dos LLMs e aquelas atribuídas pelos participantes seniores e especialistas. Com base na classificação de Schober, Boer e Schwarte (2018), foram observadas correlações positivas fortes para ChatGPT (r = 0,853), DeepSeek (r = 0,716), Llama (r = 0,775) e Gemini (r = 0,847), enquanto Claude apresentou correlação positiva moderada (r = 0,639). Esses resultados sugerem que a maioria dos modelos acompanha a tendência relativa das avaliações realizadas pelos profissionais mais experientes, embora continue atribuindo pontuações médias inferiores às desse grupo.

Tabela 11 – Coeficiente *Pearson* | LLMs vs. Seniores e Especialistas

	ChatGPT	**DeepSeek**	**Llama**	**Claude**	**Gemini**
$\overline{Humanos}$	0.853	0.716	0.775	0.639	0.847

5.4.1.3 Teste t

Terceiro, conduzimos testes estatísticos para avaliar a significância das diferenças usando testes t[4] de amostras emparelhadas para os cinco modelos e participantes humanos. Isso mostra novamente um valor p muito pequeno **(Todos** $p < 0.001$**)**, então podemos rejeitar com confiança a hipótese nula e assumir a hipótese alternativa: existe uma diferença significativa na classificação média de risco atribuída entre os dois grupos.

Os testes t emparelhados mostram na Tabela 12 que todos os LLMs diferem significativamente dos humanos Seniores e Especialistas **(**$p < 0.001$**)**. Embora os modelos sigam tendências amplamente semelhantes, as suas classificações médias são mais baixas, confirmando que os LLMs subestimam estatisticamente o risco em comparação com as avaliações dos humanos Seniores e Especialistas.

[3] Ver Schober, Boer e Schwarte (2018), para discussão sobre uso e interpretação de coeficientes de correlação.

[4] Ver *Paired Sample t-test*, Talikan et al. (2025)

Tabela 12 – Teste t emparelhado | LLMs vs. Seniores e Especialistas

Modelo	**Valor t**	**Valor p (bicaudal)**
ChatGPT	7.96	3.90×10^{-7}
DeepSeek	14.65	4.50×10^{-11}
Llama	17.03	4.06×10^{-12}
Claude	12.97	3.01×10^{-10}
Gemini	12.03	9.64×10^{-10}

5.4.1.4 Distribuição da frequência

Finalmente, analisamos os histogramas para comparar a distribuição de frequência das classificações de risco atribuídas pelos cinco modelos com aquelas fornecidas pelos participantes humanos Seniores e Especialistas. O histograma das classificações dos humanos Seniores e Especialistas na Figura 10 mostra uma clara concentração de pontuações de risco entre 3 e 6, indicando uma percepção consistente de risco maior na maioria dos cenários e muito poucas classificações abaixo de 3. Em contraste com os histogramas dos LLMs — tipicamente inclinados para valores de risco menores — a distribuição dos especialistas mostra uma tendência central mais elevada (média aritmética de ≈ 3.87) e variabilidade reduzida em valores extremos. Estas diferenças destacam a tendência dos participantes humanos em adotar uma postura mais cautelosa ao avaliar riscos potenciais, reforçando o fosso estatístico observado entre as saídas dos modelos automatizados e o julgamento profissional.

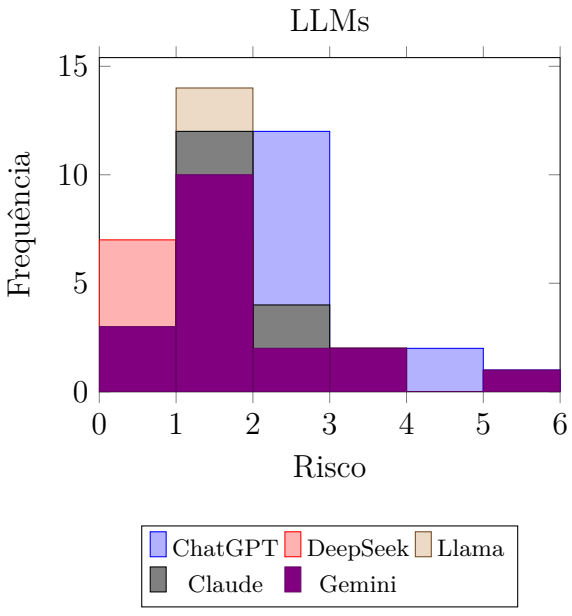

Figura 8 – Histograma de riscos | LLMs

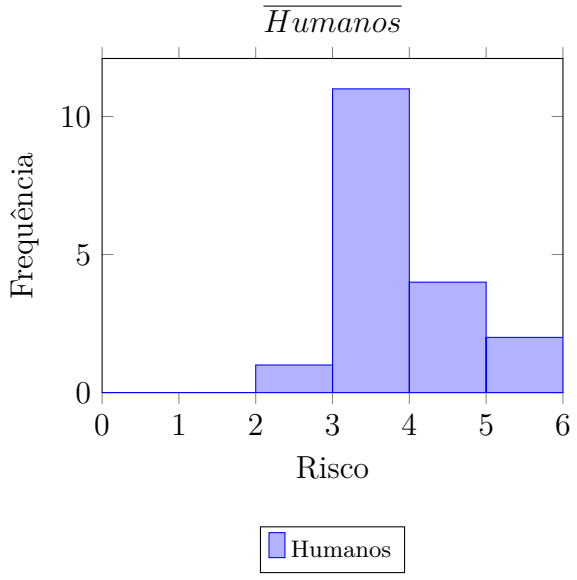

Figura 9 – Histograma de riscos | Seniores e Especialistas

Figura 10 – Histograma de riscos atribuídos | LLMs vs. Seniores e Especialistas

5.4.1.5 Considerações finais da pergunta de pesquisa 2.1

Em contraste, as distribuições dos cinco modelos permanecem fortemente concentradas nos níveis mais baixos, com apenas algumas classificações acima de 3. Estes resultados destacam como os cinco modelos continuam a subestimar o risco em relação às avaliações mais amplas e variadas feitas pelos respondentes humanos.

Nossos achados apoiam a conclusão de que os cinco modelos continuam a subestimar estatisticamente o risco de segurança da informação em relação às percepções dos participantes humanos. Esta diferença é ainda mais pronunciada quando comparada explicitamente com o grupo Sênior e Especialista, que atribuiu pontuações de risco mais elevadas do que os modelos e demonstrou um melhor acordo entre si, conforme refletido no menor desvio padrão nas suas respostas.

5.4.2 PP 2.2: LLMs vs. Humanos (Outros níveis profissionais)

Como o desempenho dos LLMs se compara aos de outros profissionais, incluindo os participantes juniores, de nível intermediário e não especialistas?

5.4.2.1 Visão geral dos riscos

Primeiro, analisamos os dados comparando o risco atribuído pelos cinco modelos com aquele atribuído pelos humanos, desta vez apenas o grupo de outros níveis profissionais (participantes que não pertencem aos grupos Sênior e Especialista). Para cada questão do cenário, calculamos o risco médio atribuído pelos cinco modelos, o risco médio atribuído pelos diferentes níveis profissionais e o desvio padrão das suas respostas.

A Tabela 13 apresenta o risco médio atribuído a cada questão pelos cinco modelos em comparação com o risco médio atribuído pelos participantes de outros níveis profissionais. O risco médio geral atribuído por este grupo de profissionais é mais alto do que o atribuído pelos cinco modelos para todas as questões, mas mais baixo do que o dos participantes humanos do grupo de Seniores e Especialistas. Além da média de risco de 3.12, observa-se um desvio padrão de 2.41, valor inferior ao observado no grupo com todos os participantes e semelhante ao observado no grupo de Seniores e Especialistas. Esse resultado sugere menor dispersão em relação ao conjunto total da amostra. A Figura 11 demonstra o gráfico de dispersão em que é possível observar como se comportam os LLMs e os humanos de outros níveis em relação aos riscos médios atribuídos para cada cenário.

Tabela 13 – Visão geral dos riscos | LLMs vs. Outros níveis profissionais

P	**ChatGPT**	**DeepSeek**	**Llama**	**Claude**	**Gemini**	$\overline{Humanos}$	**DP**
P1	1.60	0.60	1.80	1.00	0.90	3.83	3.22
P2	5.30	3.40	3.60	3.20	5.80	4.33	2.38
P3	2.60	1.60	1.80	2.00	1.60	3.11	2.35
P4	2.20	0.60	1.60	1.00	1.00	3.22	2.88
P5	2.00	0.60	1.40	1.00	0.50	2.89	2.78
P6	2.90	1.20	1.60	1.60	0.80	2.56	1.98
P7	4.60	2.40	1.80	2.00	2.10	3.39	2.40
P8	2.30	0.60	1.20	1.00	1.00	2.67	1.68
P9	2.30	1.00	1.80	1.20	1.00	2.89	2.45
P10	2.30	0.80	1.40	1.00	1.00	2.94	2.34
P11	3.60	1.80	2.20	2.60	2.20	3.17	2.55
P12	2.60	1.60	1.60	1.80	1.50	3.06	2.51
P13	2.20	0.80	1.20	1.00	1.00	2.72	2.11
P14	4.00	3.80	2.40	3.80	3.60	3.11	2.14
P15	3.80	2.60	3.40	2.80	3.10	3.28	2.30
P16	2.70	1.20	1.60	1.20	1.20	3.22	2.71
P17	2.80	1.60	1.60	1.60	1.20	3.06	2.39
P18	2.80	0.80	1.60	1.40	1.00	2.78	2.21
Total	**2.92**	**1.50**	**1.87**	**1.73**	**1.69**	**3.12**	**2.41**

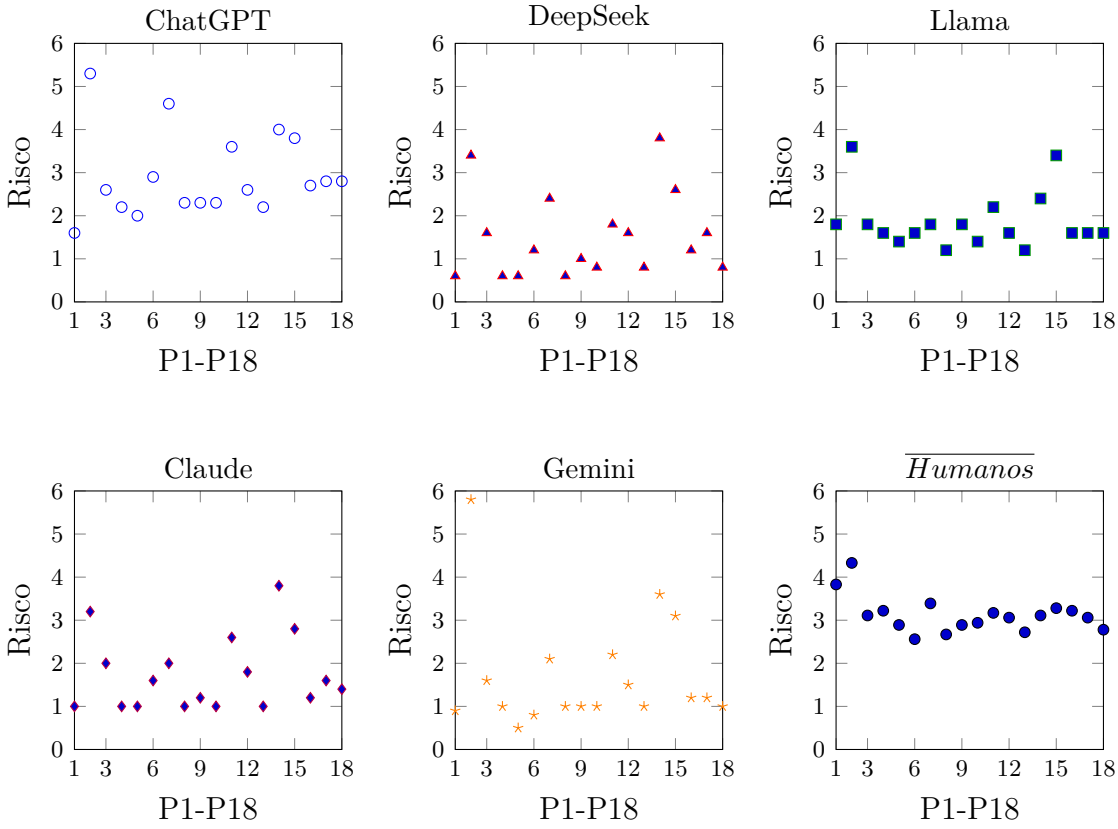

Figura 11 – Dispersão de Risco x Perguntas | LLMs vs. Outros níveis profissionais

5.4.2.2 Quantificação da associação LLMs vs. Humanos (outros níveis profissionais)

Segundo, quantificamos a relação entre os cinco modelos (ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro) e as fornecidas pelos humanos de outros níveis profissionais. Em termos práticos, existe uma correlação direta entre as classificações de risco atribuídas pelos cinco modelos e aquelas designadas pelos humanos? Ou seja, as classificações de risco movem-se na mesma direção? Para melhor compreender isso, utilizamos o coeficiente de correlação de *Pearson*[5] para medir a força da relação linear entre duas variáveis. Se houver uma forte relação linear, o coeficiente de correlação estará próximo de 1 ou -1; 0 indica que não há relação linear.

O coeficiente de correlação de *Pearson*, demonstrado na Tabela 14, indica associação positiva moderada entre as avaliações dos cinco modelos e aquelas atribuídas pelos participantes de outros níveis profissionais. Com base na classificação de Schober, Boer e Schwarte (2018), os valores observados foram classificados como moderados para ChatGPT ($r = 0{,}515$), DeepSeek ($r = 0{,}497$), Llama ($r = 0{,}688$), Claude ($r = 0{,}429$) e Gemini ($r = 0{,}698$). Esses resultados indicam que os modelos ainda acompanham parcialmente a direção das avaliações humanas, mas com menor intensidade quando comparados aos resultados observados no grupo de todos os participantes e no grupo de Seniores e Especialistas. Entre os modelos, Gemini apresentou o maior alinhamento com esse grupo, enquanto Claude apresentou o menor coeficiente de correlação.

Tabela 14 – Coeficiente *Pearson* | LLMs vs. Outros níveis profissionais

	ChatGPT	**DeepSeek**	**Llama**	**Claude**	**Gemini**
$\overline{Humanos}$	0.515	0.497	0.688	0.429	0.698

5.4.2.3 Teste t

Terceiro, conduzimos testes estatísticos conforme Tabela 15, para avaliar a significância das diferenças usando testes t[6] de amostras emparelhadas para os cinco modelos e humanos de outros níveis profissionais. Para este grupo de humanos, os resultados indicam diferenças estatisticamente significativas para DeepSeek, Llama, Claude e Gemini, todos com valores de p inferiores a 0.001. No entanto, para o ChatGPT, o valor de p foi 0.319, não permitindo rejeitar a hipótese nula. Esse resultado sugere que, no grupo de outros níveis profissionais, as avaliações médias do ChatGPT não diferiram estatisticamente das avaliações humanas, embora os demais modelos tenham mantido diferenças significativas.

[5] Ver Schober, Boer e Schwarte (2018), para discussão sobre uso e interpretação de coeficientes de correlação.

[6] Ver *Paired Sample t-test*, Talikan et al. (2025)

Tabela 15 – Teste T emparelhado | LLMs vs. Outros níveis profissionais

Modelo	Valor T	Valor P (bicaudal)
ChatGPT	1.02	3.19×10^{-1}
DeepSeek	8.11	3.00×10^{-7}
Llama	11.01	3.66×10^{-9}
Claude	7.62	6.93×10^{-7}
Gemini	5.72	2.48×10^{-5}

5.4.2.4 Distribuição da frequência

Finalmente, os histogramas demonstrados na Figura 14 mostram a distribuição das classificações de risco atribuídas pelos cinco modelos versus aquelas dadas pelos participantes humanos dos grupos de outros níveis profissionais. Este histograma mostra a distribuição da frequência das classificações de risco atribuídas pelos cinco modelos em comparação com aquelas fornecidas por este grupo de participantes. Geralmente, este grupo de participantes humanos tende a classificar entre 2 e 4, com poucas classificações acima de 3.7, indicando uma visão geral moderada do risco. Esses gráficos destacam como os modelos continuam a subestimar o risco em relação às avaliações mais amplas e variadas dos respondentes humanos.

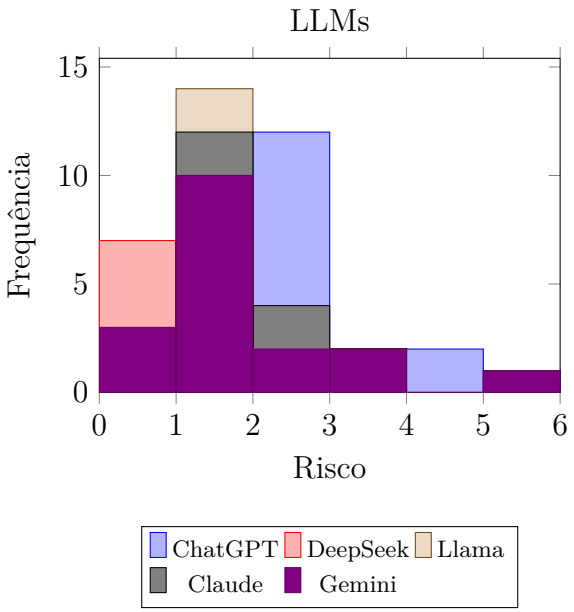

Figura 12 – Histograma de riscos | LLMs

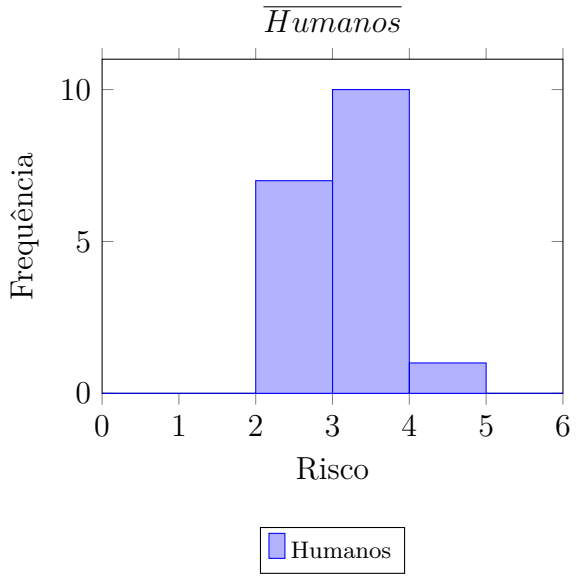

Figura 13 – Histograma de riscos | Outros níveis

Figura 14 – Histograma de riscos atribuídos | LLMs vs. Outros níveis

5.4.2.5 Considerações finais da pergunta de pesquisa 2.2

Os resultados apresentados nesta seção indicam que os modelos DeepSeek, Llama, Claude e Gemini continuam a subestimar estatisticamente os riscos de segurança da

informação quando comparados aos participantes humanos do grupo de outros níveis profissionais. Para esses modelos, as diferenças observadas foram estatisticamente significativas. O ChatGPT apresentou comportamento distinto, exibindo maior proximidade com as avaliações desse grupo e não apresentando diferença estatisticamente significativa no teste t emparelhado ($p = 0.319$). De forma geral, os resultados sugerem que os LLMs apresentam maior alinhamento com profissionais menos experientes do que com especialistas e profissionais seniores. Ainda assim, mesmo nesse grupo, a maioria dos modelos não reproduz adequadamente o julgamento humano na avaliação de riscos de segurança da informação.

5.5 Discussão e limitações

Para aumentar a clareza e facilitar a interpretação, a Tabela 16 apresenta um resumo comparativo das principais descobertas desta dissertação, que contrasta as avaliações de risco produzidas pelos cinco grandes modelos de linguagem com as dos diferentes grupos de participantes humanos, incluindo pontuações médias de risco, níveis de correlação e medidas de significância estatística. Este resumo fornece uma visão consolidada das principais divergências e padrões observados ao longo da análise. O coeficiente r representa a correlação consolidada entre a média das avaliações dos cinco LLMs e a média das avaliações humanas em cada grupo analisado.

Tabela 16 – Resumo dos resultados: LLMs vs. Humanos

Grupo	$\overline{Humanos}$	$\overline{LLMs}$	r	p
Todos os participantes	3.60	1.94	0.814	3.14×10^{-10}
Seniores & Especialistas	3.87	1.94	0.819	1.43×10^{-11}
Outros níveis	3.12	1.94	0.603	2.76×10^{-6}

Os resultados comparativos entre os três grupos indicam que os cinco modelos subestimam estatisticamente os riscos de segurança da informação em relação às avaliações humanas. Esta discrepância foi mais pronunciada entre os profissionais mais experientes (Sênior/Especialista), sugerindo que a experiência no domínio de conhecimento influencia significativamente a percepção de risco. As correlações de *Pearson* observadas, classificadas como moderadas a fortes segundo a escala de Schober, Boer e Schwarte (2018), indicam que os cinco modelos acompanham em algum grau a direção das avaliações humanas. Entretanto, embora as correlações indiquem alinhamento direcional, as diferenças absolutas e estatisticamente significativas nas pontuações demonstram que esse alinhamento não implica equivalência de julgamento.

As principais lições aprendidas incluem:

- ❏ Os LLMs seguem as tendências humanas gerais, mas falham em igualar a sensibilidade de nível de especialista a riscos com nuances;
- ❏ A experiência profissional é importante: Indivíduos mais experientes tendem a classificar o risco de forma mais elevada e consistente;
- ❏ A supervisão humana é crucial, especialmente na tomada de decisões de alto risco que envolvem informações de segurança incompletas ou ambíguas.

Estes achados sugerem que os LLMs necessitam de supervisão humana nas avaliações de risco de segurança da informação, especialmente para decisões críticas. Portanto, podem servir como ferramentas de apoio à decisão fornecendo avaliações preliminares que requerem validação. As organizações podem se beneficiar de modelos híbridos, nos quais os LLMs ajudam a padronizar as entradas ou acelerar as avaliações, enquanto os julgamentos finais permanecem com profissionais qualificados.

Além disso, as avaliações baseadas em IA podem ser mais confiáveis em contextos de ameaças externas, onde o risco é mais claramente definido e estruturado. No entanto, as avaliações de processos internos (como inventário de ativos ou adesão às políticas) revelaram as discrepâncias mais significativas e requerem uma interpretação contextual mais profunda.

Embora esta dissertação revele tendências e descobertas importantes, algumas limitações devem ser reconhecidas:

- ❏ A análise focou-se em apenas cinco modelos LLMs e pode não refletir o desempenho em outros modelos ou versões futuras;
- ❏ A amostra de participantes, embora diversificada, foi limitada a 50 indivíduos, o que pode afetar a generalização;
- ❏ O trabalho não busca eliminar as variabilidades das avaliações humanas mas utilizá-las como referência para comparação com os modelos LLMs;
- ❏ A avaliação de risco baseou-se em cenários simulados, que, embora fundamentados em práticas reais, podem não refletir totalmente a complexidade operacional;
- ❏ A metodologia desta dissertação considerou os 18 cenários de risco como observações independentes para fins analíticos. No entanto, reconhece-se a possibilidade de dependência conceitual entre eles, uma vez que todos foram derivados dos controles do CIS. Dessa forma, os resultados estatísticos devem ser interpretados como evidências comparativas dentro do conjunto experimental estudado, e não como estimativas generalizáveis para todo o universo de cenários de segurança da informação.
- ❏ Os LLMs foram testados sem engenharia de *Prompt* aprimorada ou contexto adicional que pudesse melhorar o desempenho.

Trabalhos futuros devem explorar a otimização de *Prompt*, o ajuste fino (*fine-tuning*) de modelos e o escalonamento da complexidade de cenários para avaliar como o desempenho dos LLMs pode ser melhorado em fluxos de trabalho de segurança da informação do mundo real.

Capítulo 6

Conclusão

Esta dissertação teve como objetivo investigar o uso de Grandes Modelos de Linguagem na avaliação de riscos de segurança da informação, analisando em que medida esses modelos são capazes de reproduzir julgamentos de risco realizados por profissionais humanos. Para isso, foi desenvolvido um método experimental baseado em 18 cenários de risco fundamentados nos CIS Critical Security Controls, permitindo a comparação estatística entre as avaliações produzidas por cinco LLMs amplamente utilizados e aquelas realizadas por profissionais de tecnologia e segurança da informação. A pesquisa foi orientada por duas perguntas centrais. A primeira buscou avaliar em que medida os LLMs podem ser considerados confiáveis para realizar avaliações de risco em segurança da informação em comparação com profissionais humanos. A segunda investigou como os modelos se comparam aos profissionais humanos – de acordo com sua experiência profissional – na avaliação dos cenários propostos. Para responder a essas questões, foi construído um *benchmark* contendo 18 cenários de risco, aplicado a 50 participantes humanos e aos modelos ChatGPT 5, DeepSeek V3.1, Llama 4 Scout, Claude Opus 4.1 e Gemini 2.5 Pro. As respostas obtidas foram submetidas a análises estatísticas com o objetivo de identificar padrões de convergência, divergência e significância entre os julgamentos produzidos por humanos e modelos de linguagem. Os resultados obtidos permitem responder às perguntas de pesquisa propostas e oferecem evidências relevantes sobre o potencial e as limitações dos LLMs no contexto da avaliação de riscos de segurança da informação.

Os resultados deste estudo, que avalia a confiabilidade comparativa de julgamento de riscos, apontam as diferenças entre as avaliações de risco realizadas por Humanos e aquelas geradas por grandes modelos de linguagem, LLMs. De modo geral, os modelos atribuem níveis de risco mais baixos do que os humanos, sugerindo uma subestimação do risco nos cenários analisados. Essa discrepância mostrou-se estatisticamente significativa, indicando que os modelos apresentam uma tendência de subestimação dos riscos nos cenários analisados, especialmente quando os humanos percebem alto risco. Portanto, a IA deve ser considerada uma ferramenta de apoio, e não um substituto da expertise humana.

Além disso, observamos que a experiência dos participantes influencia a atribuição de risco: profissionais mais experientes e especializados tendem a atribuir classificações de risco mais elevadas do que aqueles com menos anos de experiência. Esse fator reforça a importância da formação profissional na avaliação de riscos e sugere que os modelos de IA podem necessitar de ajustes para capturar melhor as nuances da percepção humana em contextos de segurança da informação.

Outro achado relevante foi a menor diferença entre as avaliações de risco realizadas por humanos e por LLMs em relação a ameaças externas (como riscos de terceiros). Isso pode indicar que os modelos compreendem melhor ameaças externas do que internas. Por outro lado, as maiores discrepâncias ocorreram em aspectos como inventário de ativos, nos quais especialistas humanos demonstraram maior preocupação com coberturas parciais ou incompletas, enquanto os modelos minimizaram esse risco.

Os resultados indicam que os LLMs apresentam potencial para apoiar etapas iniciais de processos de avaliação de riscos, especialmente em atividades que demandam organização, padronização e análise preliminar de informações. No entanto, os achados também evidenciam que esse apoio não deve ser confundido com equivalência ao julgamento humano especializado. Embora os modelos tenham apresentado correlação positiva com as avaliações dos participantes, as diferenças observadas nas pontuações atribuídas demonstram que acompanhar a tendência geral das avaliações humanas não significa reproduzir sua interpretação de risco. Dessa forma, a avaliação de riscos em segurança da informação permanece dependente de conhecimento contextual, experiência profissional e compreensão das consequências organizacionais associadas a cada cenário. Em especial, situações que envolvem incerteza, informações incompletas ou impactos relevantes para o negócio exigem validação humana. Assim, os LLMs devem ser compreendidos como ferramentas de apoio à decisão, capazes de auxiliar na estruturação e na análise inicial dos cenários, mas não como substitutos autônomos de especialistas humanos em processos críticos de avaliação de riscos.

6.1 Contribuições técnicas

Esta dissertação produziu contribuições técnicas relacionadas à aplicação de Grandes Modelos de Linguagem (LLMs) em processos de análise de riscos em segurança da informação. Como primeira contribuição, foi desenvolvida uma metodologia experimental para comparar avaliações de risco realizadas por participantes humanos e por múltiplos LLMs em cenários padronizados de segurança da informação.

Uma segunda contribuição consiste na construção de um conjunto estruturado de dezoito cenários de risco baseados nos controles do CIS. Os cenários foram elaborados a partir de perguntas e respostas simuladas inspiradas em processos reais de avaliação de segurança de terceiros, permitindo a realização de experimentos reproduzíveis e comparáveis

entre diferentes participantes humanos e modelos de linguagem.

Como terceira contribuição, foi criado um conjunto de dados contendo as avaliações realizadas por cinquenta participantes humanos com diferentes níveis de experiência profissional, bem como as respostas produzidas por cinco modelos de linguagem amplamente utilizados. Esse conjunto de dados possibilitou a aplicação de análises estatísticas, incluindo correlação de *Pearson*, testes t emparelhados e análises de distribuição de frequência.

As três contribuições, em conjunto, formam um *Benchmark* para futuras avaliações de riscos de segurança.

Por fim, os artefatos[1] produzidos durante a pesquisa, incluindo cenários, dados experimentais e código-fonte utilizado para análise dos resultados, foram disponibilizados publicamente, permitindo a reprodução dos experimentos e sua utilização em trabalhos futuros.

6.2 Contribuições acadêmicas e científicas

Além das contribuições técnicas, esta dissertação também gerou resultados de natureza acadêmica e científica. Como resultado direto desta pesquisa, foi elaborado e submetido um artigo científico para o periódico internacional *International Journal of Data Science and Analytics*, publicado pela *Springer Nature* e classificado como *Qualis A4*. O artigo apresenta os principais resultados desta dissertação, incluindo a metodologia experimental proposta, a comparação entre especialistas humanos e múltiplos LLMs, bem como a análise estatística dos resultados obtidos. Com o objetivo de ampliar a disseminação dos resultados para a comunidade científica, uma versão[2] preliminar do artigo foi disponibilizada publicamente no repositório arXiv.

Também foi elaborado um novo artigo científico, intitulado "Avaliação de Risco usando Engenharia de *Prompt* em LLMs", que foi aceito para publicação no SBSeg 2026. Esse trabalho complementa os resultados da dissertação ao analisar como diferentes estratégias de *prompt* podem melhorar as avaliações de risco feitas por LLMs. O artigo compara as respostas dos modelos com avaliações realizadas por profissionais da área e mostra que prompts mais direcionados, principalmente aqueles focados em evidências, lacunas e auditabilidade, ajudam a aproximar os resultados das avaliações humanas. Os LLMs têm sido explorados como ferramentas de apoio à avaliação de risco em cibersegurança, mas evidências anteriores indicam que tendem a subestimar riscos em comparação com profissionais da área. Este artigo avalia se estratégias de engenharia de *prompt* melhoram o alinhamento entre avaliações de risco produzidas por LLMs e por humanos. Os resultados mostram que *prompts* focados em análise de evidências, lacunas, cobertura parcial e auditabilidade reduziram a lacuna entre LLMs e humanos, com a melhor estratégia

[1] Ver *GitHub*: <https://github.com/gusberto/cybersecurity_assessment_llm_hitl/blob/main/>
[2] Ver <https://arxiv.org/abs/2605.05424>

reduzindo o Mean Absolute Error (MAE) em 39,2% em relação a profissionais seniores e especialistas. Ainda assim, os LLMs continuaram atribuindo notas inferiores às humanas, reforçando seu uso como ferramentas de apoio, e não como avaliadores autônomos de risco.

A publicação e disseminação desses resultados contribuem para o avanço das pesquisas relacionadas ao uso de Inteligência Artificial em segurança da informação, além de fornecer subsídios para futuras investigações sobre avaliação de riscos, interação humano-máquina e aplicação de Grandes Modelos de Linguagem em contextos organizacionais.

6.3 Trabalhos Futuros

Com base nesses achados, propomos algumas direções para trabalhos futuros:

- **Expansão do conjunto de dados:** Conduzir novos experimentos com uma amostra mais ampla e diversificada de profissionais, incluindo diferentes setores da indústria, a fim de validar e refinar os resultados;
- **Contextualização avançada:** Explorar técnicas de engenharia de *Prompts* para fornecer mais detalhes ao modelo, garantindo uma melhor compreensão do contexto das questões e avaliações mais próximas às realizadas por especialistas humanos;
- **Uso híbrido de IA e humanos:** Pesquisas futuras devem explorar abordagens híbridas nas quais sistemas de IA atuem como assistentes inteligentes de especialistas, executando tarefas preliminares como sugerir classificações iniciais, identificar padrões ou destacar riscos menos evidentes;
- **Avaliação de outros modelos de IA:** Expandir e comparar o desempenho de diferentes LLMs na avaliação de riscos em cibersegurança, investigando quais arquiteturas são mais adequadas;
- **Treinamento de modelos específicos:** Treinar e avaliar o desempenho de modelos treinados especificamente em segurança da informação, para tarefas amplas e complexas como a avaliação de riscos em diferentes cenários e contextos.

Por meio dessas iniciativas, nosso objetivo é avançar no uso da IA na avaliação de riscos em segurança da informação, tornando essa tecnologia uma aliada sólida e confiável no apoio à tomada de decisões de segurança dentro das organizações.

Referências

AGRAWAL, G. et al. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 38, n. 21, p. 23164–23172, Mar. 2024. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/30362>.

AKHTAR, Z. B. Unveiling the evolution of generative ai (gai): a comprehensive and investigative analysis toward llm models (2021–2024) and beyond. **Journal of Electrical Systems and Information Technology**, Springer, v. 11, n. 1, p. 22, 2024. Disponível em: <https://doi.org/10.1186/s43067-024-00145-1>.

AVEN, T.; RENN, O. On risk defined as an event where the outcome is uncertain. **Journal of Risk Research**, Routledge, v. 12, n. 1, p. 1–11, 2009. Disponível em: <https://doi.org/10.1080/13669870802488883>.

BAI, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. **arXiv preprint arXiv:2204.05862**, 2022. Disponível em: <https://doi.org/10.48550/arXiv.2204.05862>.

BARRETO, F. et al. Generative artificial intelligence: Opportunities and challenges of large language models. In: BALAS, V. E.; SEMWAL, V. B.; KHANDARE, A. (Ed.). **Intelligent Computing and Networking**. Singapore: Springer Nature Singapore, 2023. p. 545–553. ISBN 978-981-99-3177-4. Disponível em: <https://doi.org/10.1007/978-981-99-3177-4_41>.

BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big? In: **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2021. (FAccT '21), p. 610–623. ISBN 9781450383097. Disponível em: <https://doi.org/10.1145/3442188.3445922>.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE Trans. Pattern Anal. Mach. Intell.**, IEEE Computer Society, USA, v. 35, n. 8, p. 1798–1828, ago. 2013. ISSN 0162-8828. Disponível em: <https://doi.org/10.1109/TPAMI.2013.50>.

BHUSAL, D. et al. **SECURE: Benchmarking Large Language Models for Cybersecurity**. 2024. Disponível em: <https://doi.org/10.48550/arXiv.2405.20441>.

BOMMASANI, R. et al. On the opportunities and risks of foundation models. 08 2021. Disponível em: <https://doi.org/10.48550/arXiv.2108.07258>.

BROWN, T. B. et al. Language models are few-shot learners. In: **Proceedings of the 34th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2020. (NIPS '20). ISBN 9781713829546. Disponível em: <https://doi.org/10.48550/arXiv.2005.14165>.

CALCULATED risk? A cybersecurity evaluation tool for SMEs. **Business Horizons**, v. 63, n. 4, p. 531–540, 2020. ISSN 0007-6813. Disponível em: <https://doi.org/10.1016/j.bushor.2020.03.010>.

Center for Internet Security. **CIS Critical Security Controls Version 8**. [S.l.], 2021. Disponível em: <https://www.cisecurity.org/controls>.

CHANDRA, K. et al. **Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians**. 2026. Disponível em: <https://doi.org/10.48550/arXiv.2602.19141>.

CHENG, M. et al. Sycophantic ai decreases prosocial intentions and promotes dependence. **Science**, v. 391, n. 6792, p. eaec8352, 2026. Disponível em: <https://doi.org/10.1126/science.aec8352>.

DeepSeek-AI. Deepseek llm: Scaling open-source language models with long-term reasoning. **arXiv preprint arXiv:2401.02954**, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2401.02954>.

EKSTEDT, M. et al. Yet another cybersecurity risk assessment framework. **International Journal of Information Security**, Springer, v. 22, n. 6, p. 1713–1729, 2023. Disponível em: <https://doi.org/10.1007/s10207-023-00713-y>.

ERICSSON, K.; KRAMPE, R.; TESCH-ROEMER, C. The role of deliberate practice in the acquisition of expert performance. **Psychological Review**, v. 100, p. 363–406, 07 1993. Disponível em: <https://doi.org/10.1037/0033-295X.100.3.363>.

FERRAG et al. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. **IEEE Access**, v. 12, p. 23733–23750, 2024. Disponível em: <https://doi.org/10.1109/ACCESS.2024.3363469>.

FERRAG, M. A. et al. Generative ai and large language models for cyber security: All insights you need. **Available at SSRN 4853709**, 2024. Disponível em: <https://doi.org/10.2139/ssrn.4853709>.

FORUM, W. E. **Strategic Cybersecurity Talent Framework 2024**. [S.l.], 2024. Disponível em: <https://www3.weforum.org/docs/WEF_Strategic_Cybersecurity_Talent_Framework_2024.pdf>.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2312.11805>.

GENNARI, J. et al. Considerations for evaluating large language models for cybersecurity tasks. 2024. Disponível em: <https://www.sei.cmu.edu/documents/5848/Considerations_for_Evaluating_Large_Language_Models_for_Cybersecurity_Tasks.pdf>.

GOODFELLOW et al. Generative adversarial nets. **ArXiv**, 06 2014. Disponível em: <https://doi.org/10.48550/arXiv.1406.2661>.

GOODFELLOW, I. et al. Generative adversarial networks. **Communications of the ACM**, ACM New York, NY, USA, v. 63, n. 11, p. 139–144, 2020. Disponível em: <https://doi.org/10.1145/3422622>.

GROŠ, S. A critical view on cis controls. **arXiv preprint arXiv:1910.01721**, 2019. Disponível em: <https://doi.org/10.48550/arXiv.1910.01721>.

HAASTRECHT, M. van et al. A threat-based cybersecurity risk assessment approach addressing sme needs. In: **Proceedings of the 16th International Conference on Availability, Reliability and Security**. New York, NY, USA: Association for Computing Machinery, 2021. (ARES '21). ISBN 9781450390514. Disponível em: <https://doi.org/10.1145/3465481.3469199>.

HADI, M. U. et al. A survey on large language models: Applications, challenges, limitations, and practical usage. **Authorea Preprints**, v. 3, 2023. Disponível em: <https://doi.org/10.36227/techrxiv.23589741.v1>.

(ISC)². **Global Cybersecurity Workforce Prepares for an AI-Driven World**. [S.l.], 2024. Disponível em: <https://edu.arrow.com/media/wtjfmszx/2024-isc2-wfs.pdf>.

(ISO), I. O. for S. **ISO/IEC 27001:2022 - Information Security, Cybersecurity and Privacy Protection**. [S.l.], 2022. Disponível em: <https://www.iso.org/standard/27001>.

ISO/IEC, I. O. for S. **ISO 31000:2018, Risk management — Guidelines**. 2018. Disponível em: <https://www.iso.org/standard/65694.html>.

JI, Z. et al. Survey of hallucination in natural language generation. **ACM Computing Surveys**, ACM New York, NY, v. 55, n. 12, p. 1–38, 2023. Disponível em: <https://doi.org/10.1145/3571730>.

KAHNEMAN, D.; TVERSKY, A. Prospect theory: An analysis of decision under risk. **Econometrica**, [Wiley, Econometric Society], v. 47, n. 2, p. 263–291, 1979. ISSN 00129682, 14680262. Disponível em: <https://doi.org/10.2307/1914185>.

KAPLAN, S.; GARRICK, B. J. On the quantitative definition of risk. **Risk Analysis**, v. 1, n. 1, p. 11–27, 1981. Disponível em: <https://doi.org/10.1111/j.1539-6924.1981.tb01350.x>.

KINGMA, D.; WELLING, M. Auto-encoding variational bayes. **ICLR**, 12 2013. Disponível em: <https://doi.org/10.48550/arXiv.1312.6114>.

LEVI, M. et al. Cyberpal.ai: Empowering llms with expert-driven cybersecurity instructions. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 39, n. 23, p. 24402–24412, Apr. 2025. Disponível em: <https://doi.org/10.1609/aaai.v39i23.34618>.

LI, Z.; DUTTA, S.; NAIK, M. Llm-assisted static analysis for detecting security vulnerabilities. **arXiv preprint arXiv:2405.17238**, 2024. Disponível em: <https://doi.org/10.48550/arXiv.2405.17238>.

MCINTOSH, T. R. et al. From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. **Computers & Security**, Elsevier, v. 144, p. 103964, 2024. Disponível em: <https://doi.org/10.1016/j.cose.2024.103964>.

MERSHA, M. et al. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. **Neurocomputing**, v. 599, p. 128111, 2024. ISSN 0925-2312. Disponível em: <https://doi.org/10.1016/j.neucom.2024.128111>.

National Institute of Standards and Technology (NIST). **Framework for Improving Critical Infrastructure Cybersecurity, Version 2.0**. [S.l.], 2024. Disponível em: <https://www.nist.gov/cyberframework>.

NAVEED, H. et al. A comprehensive overview of large language models. **arXiv preprint arXiv:2307.06435**, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2307.06435>.

NIST, G. for C. R. A. **NIST SP 800-30 Rev. 1,**. 2012. Disponível em: <https://doi.org/10.6028/NIST.SP.800-30r1>.

OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2303.08774>.

OUYANG, L. et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, 2022. Disponível em: <https://doi.org/10.52202/068431-2011>.

PANKAJAKSHAN, R. et al. **Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal**. 2024. Disponível em: <https://doi.org/10.48550/arXiv.2403.13309>.

SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation coefficients: Appropriate use and interpretation. **Anesthesia & Analgesia**, v. 126, n. 5, p. 1763–1768, 2018. Disponível em: <https://doi.org/10.1213/ANE.0000000000002864>.

SentinelOne. **Cyber Security Statistics**. 2026. <https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-statistics/>. Acesso em: 29 mar. 2026. Disponível em: <https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-statistics>.

SLOVIC, P. Perception of risk. **Science**, v. 236, p. 280–285, 05 1987. Disponível em: <https://doi.org/10.1126/science.3563507>.

SLOVIC, P. et al. Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. **Risk Analysis**, v. 24, n. 2, p. 311–322, 2004. Disponível em: <https://doi.org/10.1111/j.0272-4332.2004.00433.x>.

TALIKAN, A. et al. On paired samples t-test: Applications, examples and limitations. 03 2025. Disponível em: <https://doi.org/10.5281/zenodo.10987546>.

TIHANYI, N. et al. Cybermetric: a benchmark dataset based on retrieval-augmented generation for evaluating llms in cybersecurity knowledge. In: IEEE. **2024 IEEE International Conference on Cyber Security and Resilience (CSR)**. 2024. p. 296–302. Disponível em: <https://doi.org/10.1109/CSR61664.2024.10679494>.

TOUVRON et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2302.13971>.

TOUVRON, H. et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2307.09288>.

TVERSKY, A.; KAHNEMAN, D. Judgment under uncertainty: Heuristics and biases. **Science**, American Association for the Advancement of Science, v. 185, n. 4157, p. 1124–1131, 1974. ISSN 00368075, 10959203. Disponível em: <https://doi.org/10.1126/science.185.4157.1124>.

VASWANI, A. et al. Attention is all you need. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964. Disponível em: <https://doi.org/10.48550/arXiv.1706.03762>.

VEUTHEY, J. R. et al. **MEQA: A Meta-Evaluation Framework for Question & Answer LLM Benchmarks**. 2025. Disponível em: <https://doi.org/10.48550/arXiv.2504.14039>.

WINDER, D. **Small Business Hit Hard and Often as 352 Million Records Breached**. 2026. <https://www.forbes.com/sites/daveywinder/2026/03/13/small-business-hit-hard-and-often-as-352-million-records-breached/>. Acesso em: 29 mar. 2026.

World Economic Forum. **Global Cybersecurity Outlook 2025**. Geneva: World Economic Forum, 2025. Disponível em: <https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2025.pdf>.

_____. **The Global Risks Report 2025**. Geneva: World Economic Forum, 2025. Disponível em: <https://reports.weforum.org/docs/WEF_The_Global_Risks_Report_2025.pdf>.

YAO, Y. et al. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. **High-Confidence Computing**, v. 4, n. 2, p. 100211, 2024. ISSN 2667-2952. Disponível em: <https://doi.org/10.1016/j.hcc.2024.100211>.

ZHAO, W. X. et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2303.18223>.

Apêndices

APÊNDICE **A**

Questionário para avaliação de riscos cibernéticos

A.1 Contexto e Instruções (Leia Cuidadosamente)

1. A Eva é uma especialista em segurança cibernética que trabalha para uma empresa de Cibersegurança chamada EVA CYBERSECURITY que realiza avaliações de segurança cibernética em seus clientes. O objetivo dela é determinar o risco de segurança de cada cliente (baixo, médio ou alto). Um destes clientes está sendo avaliado quanto aos riscos de segurança cibernética existentes em seu negócio (infraestrutura, processos, sistemas e aplicações, etc.).

2. A seguir, você vai encontrar os dados coletados pela Eva neste trabalho, que são:

   - ❏ As perguntas realizadas ao cliente;
   - ❏ As respostas dadas pelo cliente.

3. O seu objetivo como respondente desta pesquisa é atribuir um risco com base em sua própria experiência, sendo aquilo que você acredita, independentemente do risco atribuído pela Eva.

A.2 Análise demográfica

1. Qual a sua área de atuação?

   a) Segurança da Informação (Defensiva)

   b) Segurança da Informação (Ofensiva)

   c) Segurança da Informação (Governança)

   d) Segurança da Informação (Gestão de Riscos)

e) Tecnologia (Infraestrutura)

f) Tecnologia (Software)

g) Tecnologia (Dados)

h) Tecnologia (Governança)

i) Outras

2. Qual o seu perfil?

a) Técnico

b) Gerencial

c) Ambos

3. Você se considera em qual nível profissional?

a) Estagiário(a)

b) Júnior

c) Pleno

d) Sênior

e) Especialista

f) Outros

4. Quanto(s) ano(s) de experiência possui nessa área?

a) $<$ 1 ano

b) 1 ano

c) 2 anos

d) 3 anos

e) 4 anos

f) $>$ 5 anos

g) $>$ 10 anos

A.3 Perguntas do questionário

1. **Tema: Inventário de ativos de hardware**

   [**Pergunta para o Cliente**] A empresa possui um inventário detalhado de ativos corporativos?

[**Resposta do Cliente**] A empresa mantém um inventário de ativos abrangente e detalhado. Este inventário abrange a vasta maioria dos ativos dentro da organização, com aproximadamente 98% de todos os ativos contabilizados e rastreados. O inventário é atualizado continuamente por meio de sistemas automatizados, garantindo que novos ativos sejam capturados e os existentes sejam atualizados em tempo real conforme as mudanças ocorrem.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 17 – Tema 1 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

2. **Tema: Inventário e controle de ativos de software**

[**Pergunta para o Cliente**] A empresa possui um inventário detalhado de ativos de software, sistemas e aplicações?

[**Resposta do Cliente**] A organização tem um inventário de software bem estabelecido e efetivamente mantido. Atualmente, aproximadamente 85% do software usado dentro da organização é inventariado, garantindo controle abrangente sobre os ativos de software. Este inventário é atualizado regularmente por meio de processos manuais, garantindo que novas instalações e atualizações sejam refletidas no sistema.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 18 – Tema 2 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

3. **Tema: Gestão das Informações**

[**Pergunta para o Cliente**] Existe algum processo formal de gerenciamento de dados em vigor?

[**Resposta do Cliente**] Há um processo formal de gerenciamento de dados em vigor dentro da organização. Os dados são classificados de acordo com sua sensibilidade. O processo de gerenciamento de dados é auditado a cada seis meses, e as permissões de acesso são revisadas regularmente.

Se preferir, comente sua decisão:[ ]

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 19 – Tema 3 - Seletor de Risco

4. **Tema: Processo de configuração segura (hardening, baselines de segurança, etc.)**

   [**Pergunta para o Cliente**] Existe um processo de configuração seguro para todos os sistemas?

   [**Resposta do Cliente**] Há um processo de configuração seguro para todos os sistemas, com base na estrutura CIS (Center for Internet Security). As configurações são padronizadas em toda a organização e são revisadas e atualizadas trimestralmente.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 20 – Tema 4 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

5. **Tema: Inventários de contas de usuário**

   [**Pergunta para o Cliente**] Existe um inventário de todas as contas de usuário para todos os sistemas no catálogo de aplicativos?

   [**Resposta do Cliente**] Há um inventário de todas as contas de usuário para todos os sistemas no catálogo de aplicativos. Esse inventário é atualizado com frequência e automaticamente. As contas são gerenciadas centralmente, e as contas não utilizadas são prontamente desativadas.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 21 – Tema 5 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

6. **Tema: Concessão de acessos em sistemas e aplicações**

   [**Pergunta para o Cliente**] Existe um processo formal para conceder acesso?

   [**Resposta do Cliente**] Há um processo formal para conceder acesso. Nesse processo, as solicitações são aprovadas com base no princípio do menor privilégio, e o acesso é revisado periodicamente para garantir que ainda seja necessário. Além

disso, os acessos concedidos são baseados em papéis ou funções específicas.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 22 – Tema 6 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

7. **Tema: Gestão de Vulnerabilidades**

   [**Pergunta para o Cliente**] Existe um processo de gerenciamento de vulnerabilidades em vigor para todos os ativos (*endpoints*, servidores, dispositivos de rede, aplicativos e sistemas, etc.)?

   [**Resposta do Cliente**] Um processo de gerenciamento de vulnerabilidades está em vigor para todos os ativos, incluindo *endpoints*, servidores, dispositivos de rede, aplicativos e sistemas. Vulnerabilidades são identificadas e priorizadas para correção. Varreduras de vulnerabilidades são realizadas mensalmente, e relatórios de gerenciamento são revisados regularmente.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 23 – Tema 7 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

8. **Tema: Gestão de *logs* e registros de atividade**

   [**Pergunta para o Cliente**] Existe um processo formal de gerenciamento de *logs* de auditoria?

   [**Resposta do Cliente**] Um processo formal de gerenciamento de log de auditoria está em vigor, com coleta automática de *logs* de todos os sistemas e armazenamento seguro em repositórios centralizados e criptografados. Os *logs* são revisados regularmente por equipes dedicadas usando ferramentas automatizadas para detectar atividades suspeitas. De acordo com a política, os *logs* de auditoria são retidos por 12 meses em um ambiente seguro, com acesso restrito a usuários autorizados.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 24 – Tema 8 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

9. **Tema: Uso de navegadores e clientes de e-mail**

   [**Pergunta para o Cliente**] Todos os navegadores e clientes de e-mail são totalmente suportados e somente versões autorizadas estão implementadas?

   [**Resposta do Cliente**] Todos os navegadores e clientes de e-mail em uso são totalmente suportados, com atualizações regulares e suporte contínuo do fornecedor. As configurações de segurança para navegadores e clientes de e-mail são aplicadas centralmente e monitoradas para garantir a conformidade. A lista de versões suportadas é revisada e atualizada trimestralmente para garantir que as versões mais recentes e seguras estejam em uso.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 25 – Tema 9 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

10. **Tema: Antimalware**

    [**Pergunta para o Cliente**] O software antimalware é implantado em todos os *endpoints*?

    [**Resposta do Cliente**] O software antimalware é implantado em todos os *endpoints* organizacionais, com varreduras de *malware* conduzidas regularmente. A eficácia do software antimalware é testada trimestralmente. O recurso antiexploração é habilitado, e os *endpoints* são gerenciados centralmente por meio de uma plataforma de segurança dedicada.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 26 – Tema 10 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

11. **Tema: Recuperação dos dados e informações**

    [**Pergunta para o Cliente**] Existe um processo de recuperação de dados em vigor?

    [**Resposta do Cliente**] Um processo de recuperação de dados está em vigor, com backups realizados regularmente e armazenados com segurança. Em caso de perda, dados críticos podem ser restaurados em até 8 horas. Os *logs* de recuperação são

mantidos e revisados periodicamente para garantir a integridade do processo.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 27 – Tema 11 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

12. **Tema: Manter a infraestrutura de redes atualizada**

    [**Pergunta para o Cliente**] A infraestrutura de redes é atualizada regularmente?

    [**Resposta do Cliente**] A infraestrutura de rede é atualizada regularmente, com *patches* de *firmware* e software aplicados prontamente. Todas as alterações de configuração em dispositivos de rede são rastreadas e documentadas. A infraestrutura de rede é auditada para conformidade semestralmente.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 28 – Tema 12 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

13. **Tema: Centralização dos eventos de segurança da informação**

    [**Pergunta para o Cliente**] Todos os eventos de segurança da informação e correlatos são centralizados?

    [**Resposta do Cliente**] Todos os alertas de eventos de segurança são centralizados em uma plataforma dedicada, permitindo o monitoramento em tempo real de toda a infraestrutura. Todos os eventos críticos são enviados para esse sistema central, onde são analisados e endereçados por equipes especializadas.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 29 – Tema 13 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

14. **Tema: Programa de conscientização**

    [**Pergunta para o Cliente**] Atualmente existe algum programa com o intuito de conscientizar ou treinar os colaboradores quanto ao tema da segurança cibernética?

[**Resposta do Cliente**] Um programa ativo de conscientização sobre segurança está em vigor, e todos os funcionários são obrigados a concluir o treinamento de conscientização sobre segurança. O treinamento é realizado anualmente.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 30 – Tema 14 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

15. **Tema: Gestão de terceiros, parceiros e fornecedores**

[**Pergunta para o Cliente**] Como é feita a gestão de terceiros, parceiros e fornecedores atualmente quanto ao mapeamento e análise de riscos de segurança da informação?

[**Resposta do Cliente**] Um inventário de todos os provedores de serviço é mantido e atualizado regularmente. Os provedores de serviço são revisados periodicamente para conformidade de segurança, e suas práticas de segurança são avaliadas anualmente.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 31 – Tema 15 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

16. **Tema: Desenvolvimento seguro de software**

[**Pergunta para o Cliente**] As práticas de desenvolvimento seguro de software estão implementadas?

[**Resposta do Cliente**] Um Processo de Desenvolvimento de Aplicativo Seguro é estabelecido, com práticas de segurança incorporadas em todos os estágios do ciclo de vida do desenvolvimento. As equipes de desenvolvimento seguem padrões de codificação seguros, e revisões de segurança regulares, como testes de vulnerabilidade e auditorias de código, são conduzidas. Atualizações de segurança e *patches* são aplicados continuamente para garantir a proteção dos aplicativos.

Se preferir, comente sua decisão:[ ]

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 32 – Tema 16 - Seletor de Risco

17. **Tema: Resposta a incidentes**

    [**Pergunta para o Cliente**] Existe um processo de resposta a incidentes?

    [**Resposta do Cliente**] Um Processo de Resposta a Incidentes está em vigor, com uma equipe designada responsável pelo tratamento de incidentes. O treinamento de tratamento de incidentes é conduzido regularmente, e todos os incidentes são documentados e revisados.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 33 – Tema 17 - Seletor de Risco

Se preferir, comente sua decisão:[ ]

18. **Tema: Testes e simulação de invasões**

    [**Pergunta para o Cliente**] São realizados testes de invasão, penetração ou simulações de ataques?

    [**Resposta do Cliente**] Um Programa de Teste de Penetração está em vigor, com testes regulares conduzidos em sistemas críticos e infraestrutura de rede. Equipes especializadas realizam esses testes para identificar e explorar vulnerabilidades, garantindo que ações corretivas sejam tomadas prontamente.

	1	2	3	4	5	6	7	8	9	10	
Baixo	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	Alto

Tabela 34 – Tema 18 - Seletor de Risco

Se preferir, comente sua decisão:[ ]