Avaliação de Risco usando Engenharia de *Prompt* em LLMs

Arthur P. Labaki[1], **Gustavo R. Pinto**[1], **Rodrigo S. Miani**[1]

[1]Faculdade de Computação (FACOM) – Universidade Federal de Uberlândia (UFU)

{arthur.labaki,gus,miani}@ufu.br

Abstract. *Large Language Models (LLMs) have been explored as support tools for cybersecurity risk assessment, but previous evidence showed that they tend to underestimate risks compared to professionals. This paper evaluates whether prompt engineering improves the alignment between LLM-generated and human risk assessments using one CIS Controls-based scenario with 18 questions, evaluations from 50 professionals, five LLM families, and 16 prompt conditions, totaling 7,200 individual assessments. The results show that prompts focused on evidence analysis, gaps, partial coverage, and auditability reduced the gap between LLMs and humans, with the best strategy reducing MAE by 38.9% for senior and specialist professionals. Nevertheless, LLMs continued to assign lower scores than humans, reinforcing their role as support tools rather than autonomous risk assessors.*

Resumo. *Modelos de Linguagem de Grande Escala (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, utilizando um cenário baseado nos CIS Controls com 18 perguntas, avaliações de 50 profissionais, cinco famílias de LLMs e 16 condições de prompt, totalizando 7.200 avaliações individuais. Os resultados mostram que prompts focados em análise de evidências, lacunas, cobertura parcial e auditabilidade reduziram o gap entre LLMs e humanos, com a melhor estratégia reduzindo o MAE em 38,9% em relação a profissionais seniores e especialistas. Ainda assim, as LLMs continuaram atribuindo notas inferiores às humanas, reforçando seu uso como ferramentas de apoio, e não como avaliadores autônomos de risco.*

1. Introdução

Modelos de Linguagem de Grande Escala, ou *Large Language Models* (LLMs), têm sido cada vez mais explorados como ferramentas de apoio em atividades de cibersegurança, incluindo análise de ameaças, interpretação de vulnerabilidades, resposta a incidentes, geração de documentação técnica e suporte a processos de governança, risco e conformidade [Xu et al. 2024]. Esse interesse é impulsionado pela capacidade desses modelos de processar grandes volumes de informação textual e pela necessidade crescente de automação em um cenário marcado por ameaças mais complexas, ambientes tecnológicos distribuídos e escassez de profissionais qualificados [ISC2 2024]. Nesse contexto, a avaliação de risco em cibersegurança desempenha papel central na tomada de decisão organizacional, pois permite priorizar controles, orientar investimentos, avaliar fornecedores e identificar lacunas que comprometem a postura de segurança

[NIST 2012, ISO/IEC 2022]. *Frameworks* como os CIS Controls oferecem uma base estruturada para esse tipo de avaliação [CIS 2024]; contudo, avaliar risco não depende apenas de verificar a existência de controles, mas também de interpretar evidências, maturidade dos processos, cobertura da implementação, auditabilidade e incertezas associadas a respostas incompletas ou ambíguas.

Em um estudo anterior, investigamos se LLMs poderiam realizar avaliações de risco em cibersegurança de forma semelhante a profissionais humanos, utilizando um cenário estruturado com base nos CIS Controls [Pinto et al. 2026]. Naquele experimento, cinco LLMs foram comparadas com avaliações de 50 profissionais de diferentes níveis de experiência. Os resultados mostraram que, embora os modelos apresentassem correlação moderada a forte com as avaliações humanas, eles subestimavam sistematicamente os riscos em termos absolutos, especialmente quando comparados a profissionais sêniores e especialistas. Esse resultado sugere que as LLMs conseguem acompanhar parcialmente a tendência relativa das avaliações humanas, mas não reproduzem a mesma sensibilidade ao risco observada em avaliadores mais experientes [Pinto et al. 2026].

Apesar desses resultados, permaneceu uma lacuna importante: não estava claro se a diferença observada decorria principalmente de limitações dos modelos ou da forma como a tarefa havia sido apresentada a eles. O estudo anterior utilizou uma estrutura fixa de *prompt*, sem explorar sistematicamente técnicas avançadas de engenharia de *prompt*. Como a formulação do *prompt* pode influenciar a interpretação das instruções, o tratamento de incertezas, a estruturação do raciocínio e a calibração das respostas, surge a hipótese de que estratégias mais explícitas e controladas possam reduzir o *gap* entre avaliações geradas por LLMs e avaliações humanas [Sahoo et al. 2024, Brown et al. 2020, Wei et al. 2022]. Além disso, como LLMs são atualizadas continuamente, é necessário separar possíveis ganhos decorrentes da evolução dos modelos daqueles efetivamente causados pela engenharia de *prompt*.

Com base nesse contexto, este trabalho avalia se estratégias de engenharia de *prompt* podem melhorar o alinhamento entre avaliações de risco produzidas por LLMs e por profissionais humanos. Para isso, reutilizamos a base experimental do estudo anterior, composta pelo mesmo cenário de avaliação de terceiros, 18 perguntas associadas aos CIS Controls, respostas simuladas, escala de risco de 0 a 10, grupos humanos de referência e famílias de LLMs avaliadas, agora em suas versões disponíveis no momento da nova coleta. A nova variável experimental é a condição de *prompt*. Ao todo, avaliamos 16 condições: o *prompt* original, usado como controle, e 15 variações de engenharia de *prompt*. Cada combinação entre modelo e *prompt* foi executada cinco vezes de forma independente, e os resultados foram comparados com as avaliações humanas por meio de média das notas, viés, erro absoluto médio, redução do *gap*, correlação de Pearson e estabilidade entre execuções. Assim, este trabalho busca responder às seguintes questões de pesquisa:

1. Estratégias de engenharia de *prompt* reduzem a diferença entre as avaliações de risco produzidas por LLMs e aquelas atribuídas por profissionais humanos?
2. As estratégias de *prompt* aproximam as LLMs do julgamento de profissionais sêniores e especialistas ou apenas elevam as notas de risco de forma genérica?
3. Quais tipos de estratégia de *prompt* apresentam maior impacto na calibração, consistência e aproximação das avaliações geradas pelas LLMs?

4. A reexecução do *prompt* original nas versões atuais dos modelos indica melhoria decorrente da evolução das LLMs, independentemente das novas estratégias de engenharia de *prompt*?

Este trabalho contribui ao apresentar uma extensão controlada de um estudo anterior sobre avaliação de risco com LLMs, investigando se a engenharia de *prompt* reduz a subestimação previamente observada. Além disso, compara diferentes famílias de estratégias de *prompt*, analisa se os ganhos aproximam as LLMs do julgamento de profissionais seniores e especialistas, e disponibiliza os artefatos experimentais como *benchmark* para estudos futuros.

2. Fundamentação Teórica

Nesta seção, são apresentados os conceitos fundamentais para este trabalho: avaliação de risco cibernético e CIS Controls (Subseção 2.1), bem como *Large Language Models* (LLMs) e engenharia de *prompt* (Subseção 2.2).

2.1. Avaliação de risco cibernético e CIS Controls

A avaliação de risco cibernético é central em processos de governança, risco e conformidade, pois apoia a identificação de ameaças, vulnerabilidades, impactos e prioridades de resposta [NIST 2012, NIST 2024]. Essa avaliação não se limita à existência de controles, exigindo também análise de maturidade, abrangência, qualidade das evidências, auditabilidade e risco residual [ISO/IEC 2022, NIST 2012].

Frameworks de segurança auxiliam esse processo ao fornecer uma estrutura comum para organizar práticas e controles. Entre eles, os *CIS Controls* representam um conjunto priorizado de salvaguardas voltadas à defesa contra ataques cibernéticos prevalentes, sendo utilizados como referência prática para avaliar a postura de segurança de organizações [CIS 2024]. A versão 8.1 é organizada em 18 controles e mantém compatibilidade com outros referenciais, como o *NIST Cybersecurity Framework* [CIS 2024, NIST 2024]. Neste trabalho, os CIS Controls são utilizados como referência padronizada para estruturar as 18 perguntas de avaliação e comparar notas de risco produzidas por humanos e por LLMs, conforme realizado no estudo anterior [Pinto et al. 2026].

2.2. LLMs e engenharia de *prompts*

Large Language Models (LLMs) têm sido aplicados a diferentes tarefas de apoio à cibersegurança, incluindo análise de vulnerabilidades, detecção de ameaças, interpretação de *logs*, resposta a incidentes e suporte à tomada de decisão [Xu et al. 2024]. Apesar desse potencial, suas respostas podem variar conforme o contexto, a formulação da tarefa e as instruções presentes no *prompt*. Em avaliações de risco, essa sensibilidade é especialmente relevante, pois pequenas mudanças na apresentação do problema podem alterar a interpretação de severidade, evidência e incerteza.

A engenharia de *prompt* busca direcionar o comportamento dos modelos por meio de instruções, exemplos e formatos de saída, sem alterar seus parâmetros internos [Sahoo et al. 2024]. Estratégias como *few-shot prompting*, raciocínio estruturado, rubricas, padronização de saída, extração de evidências e revisão crítica podem influenciar a forma como os modelos interpretam e respondem a tarefas complexas [Brown et al. 2020, Wei et al. 2022, Sahoo et al. 2024]. Neste estudo, essas estratégias

são investigadas como mecanismos para calibrar avaliações de risco geradas por LLMs, verificando se instruções orientadas a evidências, lacunas, auditabilidade e critérios de pontuação reduzem a diferença em relação às avaliações humanas sem apenas aumentar artificialmente as notas atribuídas pelos modelos.

3. Trabalhos Relacionados

O uso de LLMs em cibersegurança tem sido investigado em tarefas como análise de vulnerabilidades, interpretação de ameaças, geração de relatórios, resposta a incidentes e apoio à tomada de decisão. Revisões recentes destacam o potencial das LLMs para apoiar atividades de segurança, mas também apontam limitações como alucinações, dependência do contexto, inconsistência e necessidade de validação humana [Xu et al. 2024]. *Benchmarks* como SECURE, CyberMetric e MEQA avaliam LLMs em tarefas de cibersegurança, conhecimento técnico e perguntas e respostas, mas não se concentram em avaliações de risco no estilo GRC [Bhusal et al. 2024, Tihanyi et al. 2024, Veuthey et al. 2025].

Outra linha de pesquisa explora engenharia de *prompt* para melhorar o desempenho de LLMs em tarefas específicas. Estratégias como uso de exemplos, decomposição do raciocínio, definição de rubricas e padronização da saída podem alterar significativamente o comportamento dos modelos [Sahoo et al. 2024, Brown et al. 2020, Wei et al. 2022]. Em cibersegurança, essas técnicas têm sido aplicadas principalmente a tarefas técnicas, como detecção, explicação e correção de vulnerabilidades, incluindo abordagens baseadas em *chain-of-thought* para análise de falhas de *software* [Nong et al. 2024]. No entanto, esses estudos geralmente se concentram em código, conhecimento técnico ou classificação, e não em avaliações de risco no estilo GRC.

Também há trabalhos voltados à avaliação de risco, maturidade e conformidade em cibersegurança. Benz e Chatterjee propõem uma ferramenta de avaliação para PMEs baseada no *NIST Cybersecurity Framework* [Benz and Chatterjee 2020]; Van Haastrecht et al. apresentam uma abordagem de avaliação de risco voltada às necessidades de PMEs [Van Haastrecht et al. 2021]; e Ekstedt et al. propõem o Yacraf, um *framework* quantitativo que combina modelagem de ameaças e avaliação de risco [Ekstedt et al. 2023]. Esses trabalhos reforçam a importância de estruturas sistemáticas para avaliação de risco, mas não investigam diretamente a calibração de LLMs frente ao julgamento humano em cenários baseados em controles de segurança.

Mais próximo deste trabalho, nosso estudo anterior comparou avaliações de risco produzidas por cinco LLMs com avaliações realizadas por 50 profissionais de segurança e tecnologia, utilizando um cenário estruturado a partir dos CIS Controls [Pinto et al. 2026]. Os resultados indicaram que os modelos acompanhavam parcialmente a tendência relativa das avaliações humanas, mas subestimavam sistematicamente o risco em termos absolutos, especialmente quando comparados a profissionais seniores e especialistas. Diferentemente dos trabalhos anteriores, este estudo avalia sistematicamente o impacto de 15 estratégias de engenharia de *prompt* sobre a capacidade de LLMs realizarem avaliações de risco cibernético baseadas nos CIS Controls, comparando cada estratégia com o *prompt* original reexecutado nas versões atuais dos modelos e com avaliações humanas previamente coletadas.

Além da comparação experimental, este trabalho também se diferencia por dis-

ponibilizar os artefatos utilizados na avaliação, incluindo perguntas baseadas nos *CIS Controls*, respostas simuladas, avaliações humanas e respostas produzidas por diferentes LLMs sob múltiplas condições de *prompt*. Enquanto *benchmarks* existentes em cibersegurança frequentemente se concentram em conhecimento técnico, perguntas e respostas ou tarefas de código, ainda há escassez de bases voltadas especificamente à avaliação de risco no contexto de governança, risco e conformidade. Assim, a base disponibilizada neste estudo pode apoiar comparações futuras entre modelos, estratégias de *prompt* e métodos de calibração em avaliações de risco cibernético.

4. Metodologia

Este estudo foi estruturado como uma continuação direta de uma investigação anterior sobre o uso de LLMs em avaliações de risco baseadas nos CIS Controls, na qual se observou que os modelos acompanhavam parcialmente a tendência das avaliações humanas, mas atribuíam notas sistematicamente inferiores às de profissionais humanos. A partir dessa limitação, avaliamos se estratégias de engenharia de *prompt* podem reduzir essa diferença, principalmente em relação a profissionais sêniores e especialistas. Para isso, mantivemos os principais elementos experimentais do estudo anterior, incluindo o mesmo cenário de avaliação, perguntas e respostas baseadas nos CIS Controls, escala de risco, grupos humanos de referência e lógica de comparação estatística. A principal mudança metodológica foi a aplicação de diferentes variações de *prompt*, permitindo analisar seus efeitos sobre precisão, conservadorismo e consistência das avaliações geradas pelas LLMs.

4.1. Desenho do experimento

O desenho experimental foi elaborado para isolar o efeito da engenharia de *prompt* sobre as avaliações das LLMs. Mantivemos constantes o cenário, perguntas, respostas simuladas, modelos, escala de risco e referências humanas, variando apenas a condição de *prompt*. Foram avaliadas 16 condições: o *prompt* original reexecutado como controle e 15 variações de engenharia de *prompt*. A reexecução do P00 permite comparar os resultados atuais com o baseline histórico e controlar possíveis efeitos da evolução dos modelos.

O fluxo experimental adotado pode ser resumido nas seguintes etapas:

1. **Base experimental:** reutilizamos a base do estudo anterior, composta por um cenário de avaliação de terceiros, 18 perguntas associadas aos CIS Controls e respostas simuladas do parceiro avaliado.
2. **Condições de *prompt*:** aplicamos 16 condições experimentais à mesma base: o *prompt* original como controle e 15 variações de engenharia de *prompt*.
3. **Execução dos modelos:** cada condição foi aplicada às cinco LLMs avaliadas, que atribuíram notas de risco de 0 a 10 para cada pergunta. Cada combinação modelo-*prompt* foi executada cinco vezes de forma independente.
4. **Agregação dos resultados:** as cinco execuções de cada combinação foram agregadas pela média, produzindo a nota final usada nas análises quantitativas.
5. **Comparação e análise:** os resultados das LLMs foram comparados com as avaliações humanas, considerando todos os participantes, seniores e especialistas, e demais profissionais, por meio de métricas como média das notas, viés, MAE, redução do *gap*, correlação de Pearson e estabilidade entre execuções.

4.2. Base experimental e *baseline*

A base experimental utilizada neste estudo é composta por um cenário simulado de avaliação de risco de um parceiro externo. Nesse cenário, uma empresa de médio porte, que lida com informações sensíveis e depende de parceiros em atividades críticas da cadeia de suprimentos, busca avaliar o risco associado às práticas de segurança adotadas por esse parceiro. Tanto os participantes humanos quanto as LLMs assumem o papel de analistas de segurança da informação, responsáveis por interpretar as respostas fornecidas e atribuir uma nota de risco para cada item avaliado.

A avaliação contém 18 perguntas, cada uma associada a um dos 18 controles do *CIS Controls v8*, acompanhadas por respostas simuladas da empresa avaliada. Por exemplo, no Controle 1, pergunta-se se a empresa mantém um inventário detalhado de ativos de *hardware*, cuja resposta informa 98% dos ativos inventariados e atualização automática; no Controle 2, questiona-se o inventário de *softwares*, sistemas e aplicações, com resposta indicando 85% de cobertura e atualização manual. As respostas representam um cenário plausível de avaliação de terceiros, com práticas existentes, mas também limitações, lacunas ou ambiguidades que exigem julgamento profissional. Isso permite avaliar se os modelos são sensíveis a riscos residuais, ausência de evidências, cobertura parcial e maturidade dos processos.

Todas as avaliações utilizaram uma escala numérica de 0 a 10, em que 0 representa risco baixo ou inexistente e 10 representa risco elevado. A mesma escala foi aplicada a humanos e LLMs, permitindo comparação direta entre os grupos. Os dados humanos, coletados no estudo anterior, foram mantidos como referência comparativa e incluem avaliações de 50 profissionais com diferentes níveis de experiência, áreas de atuação e perfis profissionais. Para fins de análise, esses participantes foram considerados em três grupos: todos os participantes, profissionais sêniores e especialistas, e demais profissionais. O grupo de sêniores e especialistas é particularmente relevante por representar a referência humana mais próxima de uma avaliação especializada de risco em cibersegurança.

Este estudo utiliza dois *baselines*. O primeiro é o *baseline* original, correspondente aos resultados obtidos no estudo anterior com o *prompt* original, funcionando como referência histórica. O segundo é o *baseline* reexecutado, obtido pela reaplicação do mesmo *prompt* nas versões atuais dos modelos, funcionando como referência contemporânea para distinguir efeitos da evolução das LLMs daqueles produzidos pelas novas estratégias de *prompt*. Nas análises de melhoria atribuída à engenharia de *prompt*, o *baseline* reexecutado foi adotado como principal referência, pois utiliza as mesmas versões dos modelos avaliadas nas variações P01–P15, enquanto o *baseline* original foi mantido apenas como referência histórica.

4.3. Modelos avaliados e variações de *prompt*

Foram avaliadas cinco famílias de modelos equivalentes às consideradas no estudo anterior, utilizando as versões disponíveis no momento da nova coleta experimental: ChatGPT 5.5 Pro, DeepSeek V3.2, Llama 4 Scout, Claude Opus 4.7 e Gemini 3.1 Pro. Essa escolha busca preservar a comparabilidade com o estudo anterior, ao mesmo tempo em que permite observar possíveis efeitos decorrentes da evolução dos modelos. Todas as LLMs receberam o mesmo conjunto de informações: o contexto geral do cenário, as 18 perguntas

associadas aos CIS Controls e as respostas simuladas fornecidas pelo parceiro avaliado.

As 16 condições de *prompt* foram organizadas em um *prompt* de controle e 15 variações experimentais. O *prompt* de controle corresponde ao *prompt* original utilizado no estudo anterior. As demais variações foram selecionadas para cobrir famílias complementares de engenharia de *prompt* descritas na literatura, incluindo estruturação de instruções e formato de saída, rubricas de avaliação, uso de exemplos, decomposição do raciocínio, abstração de princípios, extração de evidências, revisão crítica e avaliação adversarial [Brown et al. 2020, Wei et al. 2022, White et al. 2023, Sahoo et al. 2024, Zhou et al. 2022, Zheng et al. 2024, Madaan et al. 2023, Perez et al. 2022]. A escolha dessas estratégias buscou testar mecanismos distintos que poderiam afetar a calibração das notas de risco, como explicitar critérios de pontuação, induzir raciocínio intermediário, forçar a identificação de evidências e lacunas, ou revisar criticamente a nota antes da resposta final. Todas as variações foram adaptadas ao contexto de avaliação de risco baseada nos *CIS Controls*, sem alterar o cenário, as perguntas, as respostas simuladas ou fornecer às LLMs qualquer informação sobre as notas humanas.

As 15 variações foram definidas *a priori*, antes das execuções experimentais. Para cada família identificada na literatura, foi elaborada uma versão adaptada à tarefa, com formato de saída comparável entre as condições. Nenhuma variação foi criada, ajustada ou descartada após a análise dos resultados, reduzindo o risco de viés de seleção. Embora as estratégias de *prompt* variem em sua formulação, todas seguem três princípios metodológicos. Primeiro, nenhuma variação altera o conteúdo factual do cenário, das perguntas ou das respostas. Segundo, nenhuma variação fornece às LLMs as notas atribuídas pelos participantes humanos, evitando contaminação da comparação. Terceiro, todas as variações mantêm a mesma saída principal esperada: uma nota numérica de risco para cada uma das 18 perguntas.

A Tabela 1 documenta as 16 condições avaliadas e permite analisar a engenharia de *prompt* não como uma intervenção única, mas como um conjunto de estratégias com diferentes mecanismos de orientação. Os *prompts* completos utilizados em cada condição experimental, juntamente com as respostas brutas das LLMs, *scripts* de extração e arquivos agregados de resultados, estão disponíveis em um repositório[1], permitindo a inspeção das condições P00–P15 e a reprodução das análises.

4.4. Execução dos testes e métricas de comparação

Cada combinação entre modelo e estratégia de *prompt* foi executada cinco vezes de forma independente, com o objetivo de reduzir variações ocasionais nas respostas. Considerando cinco modelos, 16 condições de *prompt* e cinco execuções, o experimento resultou em 400 execuções completas e 7.200 avaliações individuais de risco. Em cada execução, a LLM recebeu o contexto, as 18 perguntas e as respostas simuladas, devendo atribuir uma nota de risco de 0 a 10 para cada item. Quando havia justificativas textuais, apenas a nota numérica foi usada na análise quantitativa.

Para agregar os resultados das cinco execuções, calculou-se a média das notas atribuídas por cada modelo em cada combinação de *prompt* e pergunta. Formalmente, seja $x_{m,p,q,r}$ a nota atribuída pelo modelo m, utilizando o *prompt* p, para a pergunta q, na

[1] `https://github.com/ArthurLabaki/prompt-engineering-cyber-risk-m`

Tabela 1. Resumo das estratégia de *prompt* avaliadas.

Código	Estratégia	Ideia principal
P00	*Prompt* original	Reaplica o *prompt* do estudo anterior como condição de controle.
P01	*Baseline* estruturada	Padroniza o *prompt* com papel definido, instruções claras, segmentação e saída estruturada.
P02	Rubrica ancorada simples	Define faixas explícitas para calibrar a escala de risco.
P03	Rubrica por critérios	Avalia evidência, formalização, cobertura e auditabilidade antes da nota.
P04	Penalização por falta de evidência	Trata respostas vagas, incompletas ou não auditáveis como fatores de risco.
P05	Confiança da nota	Separa a nota de risco do nível de confiança da avaliação.
P06	*Few-shot* balanceado	Usa exemplos sintéticos de diferentes níveis de risco para orientar a pontuação.
P07	*Few-shot* padrão sênior	Usa exemplos mais rigorosos para aproximar o modelo do julgamento especialista.
P08	*Least-to-Most*	Divide a avaliação em etapas menores antes da nota final.
P09	*Step-Back*	Explicita princípios gerais de risco antes de avaliar o caso concreto.
P10	*Extract-Only then Evaluate*	Primeiro extrai evidências e lacunas; depois pontua com base nelas.
P11	*Extract + Rubric*	Combina extração de evidências com rubrica de pontuação.
P12	*Self-Critique-and-Revise*	Gera uma avaliação inicial, revisa criticamente e entrega uma resposta final.
P13	*Red-Team the Score*	Testa argumentos para aumentar ou reduzir a nota antes da decisão final.
P14	Extração + revisão adversarial	Combina extração de evidências com contestação crítica da pontuação.
P15	*Ranking* primeiro, nota depois	Ordena os controles por risco relativo e depois converte o *ranking* em notas.

execução r. A média agregada para cada combinação *modelo-prompt-pergunta* é definida como:

$$\bar{x}_{m,p,q} = \frac{1}{R} \sum_{r=1}^{R} x_{m,p,q,r} \tag{1}$$

em que $R = 5$ representa o número de execuções independentes. Essa média foi utilizada como valor principal para as comparações com os grupos humanos, enquanto a variabilidade entre as cinco execuções foi analisada separadamente como medida de estabilidade.

Para cada combinação entre modelo e *prompt*, os resultados foram comparados com três referências humanas: todos os participantes, profissionais seniores e especialistas, e demais profissionais. O critério principal de melhoria foi a redução do MAE em relação aos seniores e especialistas, enquanto as demais métricas foram usadas de forma complementar. A média das notas foi analisada apenas como indicador auxiliar, pois uma estratégia poderia elevar todas as pontuações sem aproximar efetivamente o modelo do julgamento humano.

A segunda métrica foi o viés médio (*Bias*), calculado como a diferença média entre as notas das LLMs e as notas humanas. Essa diferença foi analisada tanto de forma sinalizada quanto em valor absoluto. A diferença sinalizada indica viés de subestimação ou superestimação. Valores negativos indicam que o modelo atribuiu notas inferiores

às humanas, enquanto valores positivos indicam que o modelo atribuiu notas superiores. Para um grupo humano g, a diferença sinalizada pode ser representada por:

$$Bias_{m,p,g} = \frac{1}{Q}\sum_{q=1}^{Q}(\bar{x}_{m,p,q} - h_{g,q}) \tag{2}$$

em que $Q = 18$ é o número de perguntas e $h_{g,q}$ representa a média humana do grupo g para a pergunta q.

O erro absoluto médio (*Mean Absolute Error*, ou MAE) foi utilizada para medir a distância geral entre as avaliações das LLMs e as avaliações humanas, independentemente da direção do erro:

$$MAE_{m,p,g} = \frac{1}{Q}\sum_{q=1}^{Q}|\bar{x}_{m,p,q} - h_{g,q}| \tag{3}$$

Essa métrica é importante porque permite identificar *prompts* que reduzem a distância em relação aos humanos sem depender apenas da média geral. Assim, um *prompt* é considerado mais próximo do julgamento humano quando reduz o erro absoluto médio em relação ao *baseline*.

A terceira métrica foi a redução percentual do erro (*Gap Reduction*), em relação ao *baseline* reexecutado. Essa métrica permite avaliar o ganho relativo obtido por cada técnica de engenharia de *prompt* quando comparada ao *prompt* original aplicado nas mesmas condições atuais. A redução percentual do *gap* é definida como:

$$Reduction_{m,p,g} = \frac{MAE_{m,P00,g} - MAE_{m,p,g}}{MAE_{m,P00,g}} \times 100 \tag{4}$$

em que $P00$ representa o *prompt* original reexecutado. Valores positivos indicam que a estratégia de *prompt* reduziu a distância em relação ao grupo humano analisado. Valores negativos indicam que a estratégia aumentou essa distância.

A quarta métrica foi a correlação de Pearson (*Pearson Correlation*) entre as avaliações das LLMs e as avaliações humanas. Essa métrica permite verificar se as LLMs e os humanos ordenam os cenários de forma semelhante, mesmo quando as notas absolutas são diferentes. Assim, uma estratégia de *prompt* pode ser considerada útil caso aumente a correlação com os humanos, indicando que o modelo passou a reconhecer de forma mais semelhante quais controles apresentam maior ou menor risco relativo.

A consistência entre execuções foi a estabilidade entre execuções (*Run-to-Run Stability*), medida pelo desvio padrão das notas em cada combinação *modelo-prompt-pergunta*. Essa métrica indica se uma estratégia, além de aproximar as notas das avaliações humanas, torna o comportamento do modelo mais estável e previsível. Assim, uma estratégia é interpretada como melhoria principalmente quando reduz o MAE em relação aos seniores e especialistas. Pearson e estabilidade são analisados como métricas complementares para identificar trocas entre calibração absoluta, ordenação relativa dos riscos e consistência das respostas.

5. Resultados e Discussão

Esta seção apresenta os resultados obtidos a partir da reexecução do *prompt* original e da avaliação das 15 estratégias de engenharia de *prompt* propostas. As análises utilizam como referência os mesmos grupos humanos do estudo anterior: todos os participantes, profissionais seniores e especialistas, e os demais profissionais. Como definido na metodologia, o principal critério de melhoria é a redução do erro absoluto médio (MAE) em relação ao grupo de seniores e especialistas, sem interpretar o simples aumento das notas de risco como melhoria automática.

5.1. Comparação geral com o *baseline*

A primeira análise compara o *prompt* original reexecutado (P00) com o melhor resultado obtido entre as variações de *prompt*. A Tabela 2 apresenta a comparação para os três grupos humanos. Em todos os casos, a estratégia com menor MAE foi a P04, baseada na penalização de respostas vagas, incompletas ou sem evidência auditável.

Tabela 2. Comparação entre o P00 reexecutado e o melhor *prompt* em relação aos grupos humanos.

Grupo humano	Média hum.	Média P00	Melhor *prompt*	Melhor média	MAE P00	Melhor MAE	Redução
Todos	3,60	1,95	P04	2,70	1,65	0,97	41,5%
Seniores/Especialistas	3,87	1,95	P04	2,70	1,92	1,18	38,9%
Demais profissionais	3,12	1,95	P04	2,70	1,27	0,75	41,0%

A engenharia de *prompt* reduziu consistentemente a distância entre LLMs e humanos. Para seniores e especialistas, o MAE caiu de 1,92 no P00 para 1,18 no P04, redução de 38,9%; para todos os participantes, a redução foi de 41,5%; e, para os demais profissionais, de 41,0%. Ainda assim, as LLMs permaneceram abaixo da média humana: no melhor caso geral, a média das LLMs foi 2,70, contra 3,87 dos seniores e especialistas. Portanto, as estratégias reduziram, mas não eliminaram, a subestimação observada no estudo anterior.

Para verificar se as reduções poderiam ser atribuídas à variação aleatória, aplicamos o teste de Wilcoxon pareado aos erros absolutos por pergunta (n = 18), comparando o P04 ao P00 reexecutado e considerando a média das LLMs. O MAE foi significativamente menor nos três grupos: todos os participantes (p ¡ 0,001), seniores e especialistas (p ¡ 0,001) e demais profissionais (p = 0,003). Entre seniores e especialistas, 10 das 12 variações que reduziram o MAE também apresentaram melhora significativa em relação ao P00 (p ¡ 0,05), com exceção de P06 (p = 0,064) e P12 (p = 0,144). A reexecução do P00 nas versões atuais dos modelos não indicou melhora agregada relevante: a média geral permaneceu em 1,95. A Figura 1 mostra que houve variações individuais entre modelos, como melhora do Claude e piora do GPT, mas esses efeitos se compensaram na média.

5.2. Desempenho das variações de *prompt*

A Figura 2 apresenta o MAE médio das LLMs para cada condição de *prompt*. As barras verdes indicam estratégias que reduziram o MAE em relação ao P00 reexecutado, enquanto as barras vermelhas indicam piora. De modo geral, a maior parte das estratégias

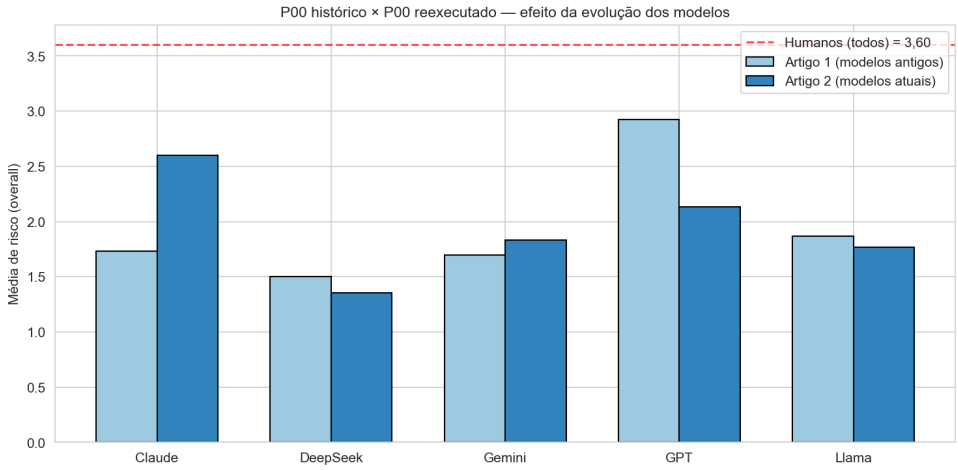

Figura 1. Comparação entre o P00 histórico e o P00 reexecutado nas versões atuais dos modelos.

melhorou o desempenho em relação ao P00. Considerando todos os participantes e o grupo de seniores e especialistas, 12 das 15 variações reduziram o MAE. Para os demais profissionais, 11 das 15 variações apresentaram melhora.

As estratégias baseadas em evidência foram as mais efetivas. Conforme definido na Tabela 1, P04 obteve o menor MAE nos três grupos humanos, seguida por abordagens como P10, P14, P15 e P11 no grupo de seniores e especialistas. Esse resultado indica que *prompts* que tornam explícita a análise de evidências, lacunas, cobertura parcial e auditabilidade tendem a aproximar mais as notas das LLMs do julgamento humano especializado

Em relação à estabilidade entre execuções, P04 também superou o P00, em que o desvio padrão médio entre execuções caiu de 0,414 para 0,312. Já P15 apresentou maior variabilidade, com desvio padrão de 0,701, indicando que elevar notas não garante consistência. Algumas estratégias não produziram ganho: P01 piorou o MAE em todos os grupos, P08 aumentou a distância em relação aos humanos e P03 foi equivalente ou ligeiramente inferior ao P00. Portanto, *prompts* mais longos ou estruturados não necessariamente produzem avaliações mais próximas das humanas.

Um ponto importante é que a melhoria não pode ser interpretada apenas pelo aumento da média das notas. A P15, por exemplo, elevou a média das LLMs para 3,19, aproximando-a da média dos demais profissionais. No entanto, essa estratégia piorou o MAE nesse grupo e apresentou menor correlação com as avaliações humanas. Isso indica que aumentar a severidade das notas de forma geral pode reduzir o viés médio, mas não necessariamente melhora a calibração pergunta a pergunta.

5.3. Comparação por modelo

A Figura 3 apresenta o MAE por combinação entre modelo e *prompt* em relação a todos os participantes. A análise indica que o efeito das estratégias não foi uniforme entre as LLMs. Algumas estratégias apresentaram melhora consistente em vários modelos, especialmente P04, P10, P11 e P14, mas a magnitude do ganho variou substancialmente.

A Tabela 3 resume, para o grupo de seniores e especialistas, o melhor *prompt* en-

Figura 2. MAE por condição de *prompt* considerando a média das LLMs.

contrado para cada modelo. Esse grupo foi utilizado por representar a principal referência humana do estudo.

Tabela 3. Melhor estratégia por modelo em relação ao grupo de seniores e especialistas.

Modelo	**MAE P00**	**Melhor *prompt***	**Melhor MAE**	**Redução**
Claude	1,36	P07	0,79	41,9%
DeepSeek	2,52	P14	1,00	60,1%
GPT	1,74	P04	0,86	50,8%
Gemini	2,04	P11	1,48	27,5%
Llama	2,11	P06	1,66	21,1%
Média das LLMs	1,92	P04	1,18	38,9%

O DeepSeek foi o modelo mais sensível às estratégias de *prompt*, com redução de 60,1% no MAE em relação aos seniores e especialistas. O GPT também apresentou ganho expressivo, com redução de 50,8%. Por outro lado, Gemini e Llama apresentaram melhorias mais moderadas. Além disso, o *prompt* com menor MAE variou entre modelos: P04 foi o melhor segundo esse critério para a média geral das LLMs e para o GPT, mas Claude teve menor MAE com P07, DeepSeek com P14, Gemini com P11 e Llama com P06. Esse resultado indica que não há uma única estratégia universalmente ótima para todas as LLMs.

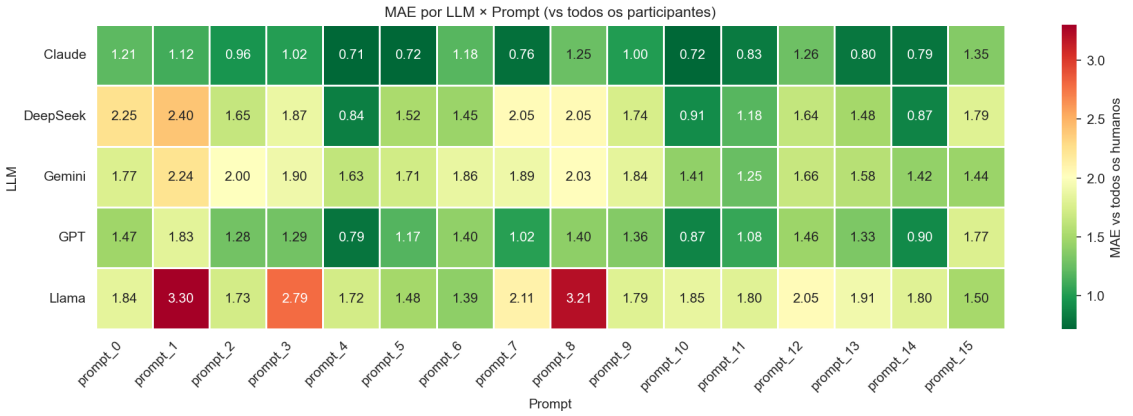

Figura 3. MAE por combinação entre LLM e *prompt* em relação a todos os participantes.

5.4. Comparação com os grupos humanos

As estratégias reduziram o MAE nos três grupos humanos, mas a melhora mais relevante ocorreu em relação aos seniores e especialistas, referência humana principal do estudo. Nesse grupo, o MAE caiu de 1,92 para 1,18, embora a média das LLMs ainda tenha permanecido inferior. Para todos os participantes, o MAE caiu de 1,65 para 0,97; para os demais profissionais, de 1,27 para 0,75. Como esse último grupo já atribuía notas mais baixas, sua maior proximidade inicial não deve ser interpretada como evidência suficiente de calibração especializada.

Também se observa uma troca entre reduzir o erro absoluto e preservar a ordenação relativa dos riscos. No grupo de seniores e especialistas, por exemplo, a correlação de Pearson caiu de 0,814 no P00 para 0,662 no P04, embora o MAE tenha sido reduzido de 1,92 para 1,18. Para todos os participantes, a correlação passou de 0,805 para 0,637; para os demais profissionais, de 0,617 para 0,430. Por outro lado, estratégias como P14 preservaram melhor a correlação relativa, embora com MAE ligeiramente maior. Isso sugere que diferentes *prompts* podem favorecer objetivos distintos: calibração absoluta das notas ou preservação do *ranking* relativo dos riscos.

5.5. Discussão

Os resultados indicam que a engenharia de *prompt* melhora a calibração de LLMs em avaliações de risco cibernético, mas não elimina a subestimação identificada no estudo anterior. Mesmo no melhor cenário, os modelos continuaram atribuindo notas médias inferiores às humanas, reforçando seu uso como ferramentas de apoio, e não como substitutos de avaliadores humanos.

As estratégias mais efetivas foram aquelas alinhadas ao domínio da tarefa, especialmente as que explicitaram evidências, lacunas, cobertura parcial, maturidade e auditabilidade. Em contraste, *prompts* apenas estruturais ou baseados em decomposição genérica do raciocínio não necessariamente melhoraram os resultados. Além disso, o impacto variou entre modelos, indicando que não há uma única estratégia universalmente ótima.

Do ponto de vista prático, os resultados sugerem que LLMs podem ser usadas em sistemas de apoio à avaliação de risco, especialmente para triagem, pré-análise e revisão

de questionários de terceiros. Um sistema desse tipo poderia extrair evidências e lacunas das respostas, apontar baixa auditabilidade, sugerir uma pontuação preliminar e destacar itens que exigem revisão especializada. Estratégias como P04, baseada em penalização por falta de evidência, e P10 e P11, baseadas em extração explícita de evidências e lacunas, são adequadas para esse cenário por tornarem a avaliação mais rastreável. No entanto, como as LLMs ainda atribuíram notas inferiores às humanas, seu uso mais seguro é como primeira camada de apoio em ferramentas de GRC, mantendo a decisão final com profissionais humanos.

Por fim, o estudo apresenta três limitações. Primeiro, utiliza um único cenário de avaliação de terceiros, com respostas simuladas, o que restringe a generalização para outros contextos de GRC. Segundo, as avaliações humanas servem como referência prática, não como verdade absoluta: o consenso de seniores e especialistas representa o julgamento especializado disponível, sem garantia de exatidão. Terceiro, a análise considera apenas as notas numéricas, e não a qualidade das justificativas textuais dos modelos, cuja avaliação fica como trabalho futuro.

6. Considerações Finais

Este trabalho investigou se estratégias de engenharia de *prompt* melhoram o alinhamento entre avaliações de risco cibernético produzidas por LLMs e por profissionais humanos. Para isso, reutilizou-se uma base baseada nos *CIS Controls*, com 18 perguntas, respostas simuladas, escala de risco de 0 a 10 e avaliações de 50 profissionais, avaliando o *prompt* original reexecutado e 15 variações de engenharia de *prompt*.

Os resultados mostram que a engenharia de *prompt* reduz de forma relevante a distância entre LLMs e humanos. A melhor estratégia reduziu o MAE em 38,9% em relação aos profissionais seniores e especialistas. As abordagens mais eficazes foram orientadas a evidências, lacunas, cobertura parcial, maturidade e auditabilidade. Ainda assim, as LLMs permaneceram abaixo das médias humanas, reforçando seu uso como apoio, e não como substitutas autônomas em processos de GRC.

Outra contribuição relevante deste trabalho é a disponibilização da base experimental utilizada, composta por perguntas associadas aos *CIS Controls*, respostas simuladas, avaliações humanas, respostas produzidas pelas LLMs, condições de *prompt* e resultados agregados. Essa base permite a reprodução dos experimentos e pode servir como *benchmark* para futuros estudos que investiguem novas LLMs, novas estratégias de engenharia de *prompt*, métodos de calibração ou abordagens híbridas de avaliação de risco envolvendo humanos e modelos generativos.

Como trabalhos futuros, pretende-se ampliar a base experimental para outros cenários, setores e *frameworks*, como NIST CSF e ISO/IEC 27001. Também será relevante investigar abordagens híbridas, nas quais LLMs produzam avaliações preliminares revisadas por especialistas humanos. Além disso, uma direção promissora é analisar qualitativamente as justificativas geradas pelos modelos, verificando se o raciocínio apresentado é consistente, auditável e adequado para apoiar decisões reais de risco cibernético.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) projetos 409743/2025-9 e 408432/2024-1.

Referências

Benz, M. and Chatterjee, D. (2020). Calculated risk? a cybersecurity evaluation tool for smes. *Business horizons*, 63(4):531–540.

Bhusal, D., Alam, M. T., Nguyen, L., Mahara, A., Lightcap, Z., Frazier, R., Fieblinger, R., Torales, G. L., Blakely, B. A., and Rastogi, N. (2024). Secure: Benchmarking large language models for cybersecurity. In *2024 Annual Computer Security Applications Conference (ACSAC)*, pages 15–30. IEEE.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

CIS (2024). CIS Critical Security Controls v8.1. White paper. Publicado em 24 jun. 2024. Acesso em: 3 maio 2026.

Ekstedt, M., Afzal, Z., Mukherjee, P., Hacks, S., and Lagerström, R. (2023). Yet another cybersecurity risk assessment framework. *International Journal of Information Security*, 22(6):1713–1729.

ISC2 (2024). Global cybersecurity workforce prepares for an ai-driven world. Technical report, ISC2.

ISO/IEC (2022). ISO/IEC 27005:2022 information security, cybersecurity and privacy protection — guidance on managing information security risks. `https://www.iso.org/standard/80585.html`.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

NIST (2012). Guide for conducting risk assessments. Technical Report NIST Special Publication 800-30 Revision 1, National Institute of Standards and Technology.

NIST (2024). The nist cybersecurity framework (csf) 2.0. Technical Report NIST CSWP 29, National Institute of Standards and Technology.

Nong, Y., Aldeen, M., Cheng, L., Hu, H., Chen, F., and Cai, H. (2024). Chain-of-thought prompting of large language models for discovering and fixing software vulnerabilities. *arXiv preprint arXiv:2402.17230*.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. (2022). Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

Pinto, G. R., do Prado Labaki, A., and Miani, R. S. (2026). Evaluating the reliability of multiple large language models in risk assessment: A cis controls based approach.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1.

Tihanyi, N., Ferrag, M. A., Jain, R., Bisztray, T., and Debbah, M. (2024). Cybermetric: A benchmark dataset based on retrieval-augmented generation for evaluating llms in

cybersecurity knowledge. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 296–302. IEEE.

Van Haastrecht, M., Sarhan, I., Shojaifar, A., Baumgartner, L., Mallouli, W., and Spruit, M. (2021). A threat-based cybersecurity risk assessment approach addressing sme needs. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–12.

Veuthey, J. R., Majid, Z. A., Hariharan, S., and Haimes, J. (2025). Meqa: A meta-evaluation framework for question & answer llm benchmarks. *arXiv preprint arXiv:2504.14039*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., and Wang, H. (2024). Large language models for cyber security: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*.

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., and Zhou, D. (2024). Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations*, volume 2024, pages 20279–20316.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.