Impacto do Uso de *Large Language Models* na Interpretação de Requisitos: Um Estudo Experimental Controlado

Gisely Meneses de Oliveira
Instituto de Informática, Universidade Federal de Goiás (UFG)
Goiânia, GO, Brasil
giselymoli@gmail.com

Valdemar Vicente Graciano Neto
Instituto de Informática, Universidade Federal de Goiás (UFG)
Goiânia, GO, Brasil
valdemarneto@ufg.br

Resumo

A Engenharia de Requisitos é uma das fases mais críticas do desenvolvimento de *software*, e a crescente adoção de *Large Language Models* (LLMs) torna necessário investigar seu impacto na interpretação de requisitos. Este artigo relata um experimento controlado intra-sujeitos (*within-subjects*) com contrabalanceamento, conduzido com 35 participantes, para avaliar o efeito do ChatGPT nessa tarefa. Cada participante executou duas tarefas equivalentes, uma com apoio do LLM e outra sem, a partir de cenários com itens plantados (requisitos, ambiguidades e omissões) validados por especialistas. Foram avaliadas cinco variáveis dependentes comparativas (completude, corretude, identificação de problemas, tempo e carga cognitiva pelo NASA-TLX) e uma medida de utilidade percebida (TAM), esta última aplicada exclusivamente na condição com LLM. Os resultados fornecem evidência de que, no contexto estudado, o uso do ChatGPT esteve associado a melhorias em todas as métricas comparativas (completude: +11,9 p.p.; corretude: +8,5 p.p.; identificação de problemas: +1,3 itens; tempo: −25,8%; carga cognitiva: −19,6%), com significância estatística ($p < 0{,}001$) e efeitos grandes, robustos à correção de Bonferroni. A utilidade percebida do LLM foi alta (M = 5,96/7,0). Discute-se também o compromisso entre o tempo economizado na geração e o tempo requerido para verificação da corretude das saídas do LLM.

Palavras-chave

Engenharia de Requisitos, *Large Language Models*, ChatGPT, Elicitação de Requisitos, Experimento Controlado, Qualidade de Software

1 Introdução

A Engenharia de Requisitos (ER) é amplamente reconhecida como uma das fases mais críticas do desenvolvimento de *software*. Falhas nessa etapa respondem por cerca de dois terços dos fracassos em projetos [1] e por 47% dos projetos malsucedidos segundo dados do PMI reportados por Vazquez e Simões [2]. O custo de corrigir defeitos cresce de forma quase exponencial à medida que o projeto avança, sendo os defeitos originados na elicitação os mais onerosos [1].

Com o amadurecimento dos *Large Language Models* (LLMs), ferramentas como o ChatGPT passaram a integrar o cotidiano de engenheiros de *software*. Estudos recentes exploram a aplicação de LLMs em atividades de ER, incluindo classificação de requisitos, detecção de ambiguidades e geração de histórias de usuário [3, 4]. Ronanki et al. [4] observaram que LLMs podem superar especialistas humanos em atributos estruturais como consistência e atomicidade, mas apresentam deficiências em não-ambiguidade e viabilidade quando operam de forma autônoma. Arora et al. [3] ressaltam que a qualidade dos resultados depende fortemente da experiência do engenheiro na formulação de *prompts*.

Entretanto, a literatura é predominantemente exploratória, com poucos estudos experimentais controlados sobre o efeito do uso de LLMs na interpretação humana de requisitos durante a elicitação. Esta lacuna motiva o estudo aqui relatado: um experimento controlado intra-sujeitos para avaliar se, e em que medida, o uso do ChatGPT, como ferramenta de apoio ao raciocínio humano e não como substituto, afeta a qualidade da interpretação de requisitos por praticantes e estudantes da área de tecnologia.

O relato segue as diretrizes para condução e reporte de experimentos em Engenharia de Software de Wohlin et al. [5] e Jedlitschka e Pfleeger [9].

2 Referencial Teórico

2.1 Engenharia De Requisitos E Elicitação

A Engenharia de Requisitos compreende as atividades voltadas à identificação, documentação e validação das necessidades das partes interessadas em relação a um sistema de *software* [5]. A elicitação, processo de descoberta e levantamento de requisitos junto aos *stakeholders* tipicamente por meio de entrevistas, é especialmente crítica, pois ambiguidades e omissões nessa fase propagam-se para todas as etapas subsequentes [1, 2]. A interpretação de requisitos, objeto deste estudo, refere-se à capacidade do engenheiro de extrair, estruturar e identificar inconsistências a partir das informações brutas fornecidas pelos *stakeholders*. Atributos como completude, corretude e não-ambiguidade são centrais para avaliar essa capacidade [4].

2.2 Métricas De Qualidade Adotadas

Este estudo operacionaliza a qualidade da interpretação de requisitos por seis variáveis dependentes. Completude e corretude capturam, respectivamente, a abrangência e a precisão na identificação de requisitos funcionais e não funcionais, ancoradas no *gold standard* de cada cenário. A identificação de problemas, definida como a contagem de ambiguidades e omissões detectadas (máx. = 8 por tarefa), é especialmente relevante, dado que LLMs autônomos tendem a apresentar deficiências nesse atributo [4]. O tempo captura o esforço cronológico. A carga cognitiva é mensurada pelo NASA-TLX [6]. A utilidade percebida é capturada por itens adaptados do TAM (*Technology Acceptance Model*) [7].

3 Definição E Planejamento Do Experimento

3.1 Definição De Objetivos (GQM)

Seguindo o *template Goal-Question-Metric* (GQM) [5]: analisar o uso de *Large Language Models*, com o propósito de avaliar seu efeito, com respeito à qualidade da interpretação de requisitos, do ponto de vista de pesquisadores e profissionais de ER, no contexto de praticantes e estudantes da área de tecnologia executando tarefas controladas de interpretação a partir de transcrições simuladas de entrevistas com *stakeholders*.

A questão de pesquisa (QP1) é: como o uso do ChatGPT afeta a qualidade da interpretação de requisitos por praticantes e estudantes durante atividades de elicitação? As hipóteses formais são: H0: o uso de LLM não altera a qualidade da interpretação de requisitos; H1: o uso de LLM altera a qualidade da interpretação de requisitos. O nível de significância adotado é $\alpha = 0{,}05$.

3.2 Seleção Da Ferramenta

A escolha do ChatGPT como instância de LLM justifica-se por quatro razões: (i) é a ferramenta predominante na literatura recente de LLMs em ER, o que permite comparabilidade direta com estudos anteriores [3, 4]; (ii) é o LLM de maior adoção entre profissionais (na própria amostra deste estudo, 82,9% relataram uso diário de LLMs, sendo o ChatGPT o mais citado); (iii) sua interface *web* gratuita reproduz a condição real de uso da maioria dos praticantes, favorecendo a validade ecológica do experimento; e (iv) a versão exata do modelo e a data de acesso foram registradas, permitindo reprodutibilidade.

Considerações éticas. A participação foi voluntária, condicionada ao Termo de Consentimento Livre e Esclarecido (TCLE), com direito de desistência sem prejuízo; dados anonimizados por identificadores individuais e reportados de forma agregada; a sessão encerrou-se com um *debriefing* formativo aos participantes.

3.3 Variáveis

A variável independente é o uso de LLM, com dois níveis (fatores): com apoio do ChatGPT e sem apoio. As seis variáveis dependentes são: (1) completude: proporção de requisitos funcionais (RF) e não funcionais (RNF) do *gold standard* identificados (máx. = 8 itens por cenário); (2) corretude: proporção de itens enunciados alinhados ao *gold standard*; (3) identificação de problemas: número de ambiguidades e omissões detectadas (máx. = 8: 4 ambiguidades + 4 omissões); (4) tempo: duração total da tarefa em minutos; (5) carga cognitiva: NASA-TLX versão curta; e (6) utilidade percebida: itens adaptados do TAM, aplicados apenas após a condição com LLM. As variáveis de controle registradas foram: anos de experiência, experiência prévia em ER, frequência de uso de LLMs e nível de formação.

3.4 Desenho Experimental E Sua Justificativa

O experimento adota desenho intra-sujeitos (*within-subjects*) com contrabalanceamento da ordem de tratamento. Cada participante executou duas tarefas equivalentes em domínios distintos: participantes da Sequência A executaram o Cenário A (sistema de biblioteca universitária) com apoio do ChatGPT e o Cenário B (aplicativo de *delivery* de refeições) sem apoio; participantes da Sequência B executaram a ordem inversa.

Uma alternativa mais simples seria o desenho entre-sujeitos (*between-subjects*), com um grupo de intervenção e um grupo de controle. O desenho intra-sujeitos foi preferido por três razões, alinhadas a diretrizes clássicas de desenho experimental [5, 10, 11]. Primeiro, o poder estatístico: para detectar um efeito médio (d = 0,5) com $\alpha = 0{,}05$ e poder de 0,80, o desenho pareado exige n ≈ 34 participantes, enquanto o desenho entre-sujeitos exigiria cerca de 63 por grupo (≈ 126 no total), o que seria inviável no contexto de recrutamento disponível [12]. Segundo, o controle da heterogeneidade individual: a experiência dos participantes variou de 1 a 26 anos; no desenho pareado, cada participante atua como seu próprio controle, eliminando essa variabilidade da comparação, ao passo que no desenho entre-sujeitos ela inflaria o erro e exigiria balanceamento cuidadoso dos grupos [10]. Terceiro, a viabilidade logística: o desenho pareado permitiu a execução em sessão única de 60 minutos.

Reconhecem-se, contudo, limitações intrínsecas ao desenho intra-sujeitos que impactam a generalização dos resultados. Efeitos de aprendizagem, fadiga e transferência entre condições são ameaças típicas desse desenho [11, 12]; foram mitigados pelo contrabalanceamento de ordem e pelo uso de cenários equivalentes em domínios distintos, mas não podem ser completamente eliminados. Adicionalmente, o desenho pareado favorece a validade interna (comparação intra-indivíduo) em detrimento da validade externa: os resultados descrevem o efeito médio dentro do mesmo indivíduo em duas condições, não o efeito populacional de adoção do LLM. Assim, generalizações para populações e contextos distintos requerem replicações com desenhos entre-sujeitos ou fatoriais mistos [5, 10].

3.5 Participantes

A população-alvo compreende praticantes e estudantes da área de tecnologia. O critério de inclusão exigiu participação em ao menos um projeto com levantamento de requisitos ou disciplina formal de ER. O tamanho amostral foi calculado a priori no G*Power para teste t pareado bilateral, com efeito médio (d = 0,5), $\alpha = 0{,}05$ e poder = 0,80, resultando em n mínimo de 34.

A amostra final (n = 35) apresenta perfil heterogêneo: idade média de 34,3 anos (18–54 anos), experiência na área de tecnologia com média de 10,7 anos (Mdn = 8; 1–26 anos), majoritariamente com pós-graduação (especialização ou mestrado) e atuando em desenvolvimento de *software*, gestão de projetos, QA/testes e funções correlatas. Todos os participantes já haviam atuado em ao menos uma atividade de levantamento de requisitos, e 22,9% haviam cursado disciplina formal de Engenharia de Requisitos. O uso de ferramentas de IA generativa era frequente, com 82,9% relatando uso diário; ChatGPT, Claude e Copilot figuravam entre as ferramentas mais utilizadas. Essa heterogeneidade em experiência, formação e papel favorece a validade externa dentro do perfil-alvo, ainda que a autosseleção (voluntários de comunidades técnicas) permaneça como fator a considerar (Seção 6.3).

3.6 Instrumentação

Cada cenário apresenta uma transcrição simulada de entrevista com *stakeholder*, de aproximadamente 350–360 palavras (2–3 min

de leitura), com estrutura idêntica: 6 requisitos funcionais, 2 não funcionais, 4 ambiguidades e 4 omissões plantadas, totalizando 16 itens avaliáveis por cenário.

Construção e validação do gold standard. A construção inicial dos cenários e do *gold standard* foi realizada pelos pesquisadores responsáveis pelo delineamento do experimento, definindo, para cada item plantado, sua categoria (RF/RNF/ambiguidade/omissão) e a evidência textual correspondente. O objetivo foi assegurar que ambos os cenários apresentassem complexidade equivalente e permitissem avaliar de forma consistente a identificação de requisitos e problemas de qualidade. Em seguida, cenários, *gold standard* e rubrica foram submetidos à validação por dois especialistas em ER, com experiência acadêmica e profissional em elicitação, especificação, validação e revisão de requisitos. Os especialistas verificaram a completude do *gold standard*, a consistência entre cenários e respostas esperadas, a existência de interpretações alternativas, a adequação das ambiguidades e omissões planejadas, e a equivalência de complexidade entre os cenários A e B. Divergências foram discutidas entre especialistas e pesquisadores até consenso: quando necessário, requisitos foram reformulados, ambiguidades refinadas e respostas equivalentes incorporadas ao *gold standard*. A versão consolidada foi utilizada como referência única durante toda a etapa de correção.

Rubrica de avaliação. A rubrica traduz cada variável dependente em critérios verificáveis. Dois avaliadores a aplicaram de forma independente e cega quanto à condição experimental, com concordância inter-avaliadores mensurada pelo κ de Cohen e critério mínimo aceitável pré-estabelecido de $\kappa \geq 0{,}70$. Nos casos de discordância, a classificação final foi definida por consenso entre os avaliadores, utilizando o *gold standard* como referência.

Prompt padronizado. Na condição com LLM, um *prompt* padronizado instruía o modelo a estruturar a resposta em quatro categorias (RF, RNF, ambiguidades e omissões) e a não inventar informações ausentes do texto, controlando a variável qualidade do *prompt* apontada por Arora et al. [3]. O texto integral do *prompt*, dos cenários, do *gold standard* e da rubrica está disponível como material suplementar (ver seção "Disponibilidade de Artefatos").

Instruções aos participantes e formulário de respostas. Além do *prompt* padronizado, um roteiro padronizado de instruções verbais foi lido antes do início da sessão: explicava o procedimento, esclarecia aspectos logísticos e reforçava que nenhuma dúvida relativa ao conteúdo dos cenários seria respondida durante a execução. Os participantes foram orientados a: ler integralmente o cenário antes de responder; registrar todas as interpretações relevantes; não solicitar esclarecimentos conceituais aos aplicadores; e explicitar hipóteses assumidas durante a interpretação. Todos os participantes receberam exatamente as mesmas instruções. As respostas foram registradas em formulário estruturado com quatro campos obrigatórios: (1) requisitos funcionais identificados, (2) requisitos não funcionais identificados, (3) ambiguidades identificadas e (4) omissões identificadas. Esse formato padroniza a coleta e facilita a aplicação da rubrica.

Tabela 1: Síntese dos resultados (médias por condição; α = 0,05).

Variável	Com LLM	Sem LLM	Dif.	Teste	p	Efeito
Completude	82,3%	70,4%	+11,9pp	Wilcoxon	<0,001	r=0,873
Corretude	84,1%	75,5%	+8,5pp	Wilcoxon	<0,001	r=0,873
Id. probl.	5,0/8	3,7/8	+1,3	Wilcoxon	<0,001	r=0,916
Tempo (min)	8,6	11,6	−3,0	t pareado	<0,001	d=−6,25
NASA-TLX	54,3	67,6	−13,2	Wilcoxon	<0,001	r=0,873
TAM (1–7)	5,96	—	vs.4,0	t (1am.)	<0,001	grande

4 Operação Do Experimento

4.1 Preparação E Seleção Da Amostra

Inicialmente, 43 indivíduos responderam ao questionário de caracterização (TCLE, perfil demográfico, experiência e familiaridade com LLMs). Após a aplicação dos critérios de inclusão e exclusão, 8 participantes foram excluídos por não atenderem ao requisito mínimo de experiência em ER. A amostra final foi composta por 35 participantes, superando o mínimo calculado (34), alocados por sorteio em duas sequências: Sequência A (n = 18) e Sequência B (n = 17). O perfil da amostra é detalhado na Seção 3.5.

4.2 Procedimento De Coleta

O experimento foi conduzido em sessão única de 60 minutos: acomodação e distribuição de identificadores A/B por sorteio prévio (5 min); boas-vindas e retomada do consentimento (3 min); instruções e distribuição da Rodada 1 (4 min); Rodada 1 (Cenário A, com cronômetro projetado, 12 min); transição (2 min); Rodada 2 (Cenário B, condição invertida, 12 min); questionário pós-experimento com NASA-TLX, itens TAM e campo aberto (7 min); e *debriefing* formativo (12 min). Dois a três aplicadores treinados circularam pela sala atendendo apenas a dúvidas logísticas; os participantes não podiam solicitar esclarecimentos conceituais, devendo expor nas respostas as suposições assumidas.

5 Análise E Interpretação Dos Dados

5.1 Procedimento De Análise

A análise descritiva reporta médias, medianas e desvios-padrão por condição. A normalidade das diferenças pareadas foi avaliada pelo teste de Shapiro-Wilk (α = 0,05): quando sustentada, aplicou-se o teste t pareado com d de Cohen para amostras pareadas; caso contrário, o Wilcoxon *signed-rank* com tamanho de efeito r. Dado o teste simultâneo de cinco hipóteses comparativas, verificou-se adicionalmente a robustez dos resultados à correção de Bonferroni (α corrigido = 0,01). Todas as análises foram realizadas em R, com os pacotes *tidyverse*, *effsize* e *irr*.

5.2 Resultados

A Tabela 1 sintetiza os resultados e a Figura 1 apresenta as distribuições por condição. Em todas as cinco variáveis comparativas, a diferença entre as condições foi estatisticamente significativa (p < 0,001) com tamanho de efeito grande, levando à rejeição de H0. Na completude, os participantes identificaram em média um requisito adicional com apoio do LLM (6,6 vs. 5,6 de 8). Na corretude, destaca-se que a condição com LLM não atingiu 100%: parte das

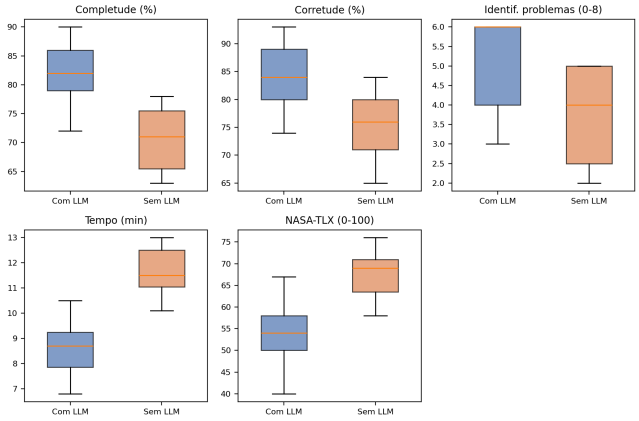

Figura 1: Distribuição das cinco variáveis comparativas por condição (n = 35).

saídas do modelo continha itens incorretos ou não ancorados no texto, que exigiram filtragem pelos participantes, aspecto retomado na Seção 6.2. Na identificação de problemas, a detecção subiu de 46,4% para 62,5% dos 8 itens plantados (Mdn: 4 vs. 6), com o maior efeito observado (r = 0,916). O tempo caiu 3,0 minutos (−25,8%). As diferenças pareadas de tempo foram a única variável com distribuição normal (Shapiro-Wilk: p = 0,072), analisada por teste t pareado; as demais, não-normais (p entre 0,013 e 0,020), foram analisadas por Wilcoxon. A carga cognitiva caiu 13,2 pontos (−19,6%). A utilidade percebida (TAM), medida apenas na condição com LLM, foi de 5,96/7,0, significativamente acima do ponto neutro de 4,0 [$t(34) = 22{,}25$; $p < 0{,}001$].

5.3 Robustez A Múltiplas Comparações

Como cinco testes de hipótese comparativos foram conduzidos, aplicou-se a correção de Bonferroni (α corrigido = 0,05/5 = 0,01). Todos os valores-p observados ($p < 0{,}001$) permanecem abaixo do limiar corrigido, indicando que os resultados são robustos ao problema de múltiplas comparações.

6 Discussão E Ameaças À Validade

6.1 Interpretação Dos Resultados

É importante delimitar o alcance das conclusões: este experimento não demonstra, de forma geral, que LLMs melhoram a interpretação de requisitos. O que os resultados fornecem é evidência empírica de que, no contexto específico estudado (participantes com o perfil descrito, tarefas curtas de interpretação a partir de transcrições simuladas e ChatGPT com *prompt* padronizado), o uso do LLM esteve associado a melhorias estatisticamente significativas e de efeito grande em todas as métricas adotadas. Generalizações para outros contextos, ferramentas e formatos de tarefa requerem replicações.

Feita essa ressalva, o padrão observado é consistente. O resultado relativo à identificação de problemas merece destaque: Ronanki et al. [4] relataram deficiência de LLMs no atributo de não-ambiguidade quando operam de forma autônoma. Neste estudo, com o LLM atuando como apoio ao raciocínio humano por meio de *prompt* estruturado, que solicita explicitamente a categorização de ambiguidades e omissões, observou-se o efeito oposto: os participantes identificaram mais problemas no texto (Mdn: 6 vs. 4 de 8; r = 0,916). Uma explicação plausível é que o *prompt* estruturado atua como uma lista de verificação cognitiva, direcionando a atenção do engenheiro a categorias que, sem o apoio, tenderiam a ser negligenciadas.

A redução simultânea de tempo e carga cognitiva, combinada com o aumento da qualidade, é compatível com a interpretação do LLM como organizador cognitivo: a ferramenta estrutura a saída em categorias, reduz o esforço de classificação e libera capacidade cognitiva para a análise crítica. A alta utilidade percebida (M = 5,96/7,0) é coerente com essa leitura e com os achados de Arora et al. [3].

Cabe reconhecer, contudo, que os ganhos observados não podem ser atribuídos exclusivamente ao LLM. O *prompt* padronizado adotado solicitava explicitamente que a resposta fosse organizada em quatro categorias (RF, RNF, ambiguidades e omissões), o que confere ao próprio *prompt* o papel de lista de verificação cognitiva (*cognitive checklist*). Parte da melhoria em identificação de problemas e completude pode, portanto, ser atribuível ao efeito estruturante do *prompt* e não ao LLM em si. Um desenho fatorial que dissociasse as duas fontes de efeito, contrastando (i) LLM sem *prompt* estruturado, (ii) *checklist* sem LLM e (iii) LLM com *prompt* estruturado, é necessário para isolar essas contribuições e configura direção prioritária de trabalho futuro.

6.2 O Compromisso Entre Geração E Verificação

Um aspecto relevante é o compromisso (*trade-off*) entre o tempo economizado na geração e o tempo requerido para verificação. A saída do LLM não é isenta de erros: pode conter interpretações equivocadas, itens não ancorados no texto e alucinações. A corretude média de 84,1% na condição com LLM, e não 100%, evidencia que parte das saídas continha itens que os participantes precisaram filtrar, corrigir ou descartar, e que essa filtragem foi imperfeita. Assim, parcela do tempo poupado na produção da lista de requisitos é reinvestida na conferência da corretude do material gerado.

Na prática, o uso do LLM de fato reduz o tempo de construção da lista de requisitos; contudo, permanece um tempo que o analista precisa dedicar à análise e ao refinamento da saída: ler o material entregue pelo modelo, realizar ajustes de redação e escopo, e até descartar informações que, embora plausíveis, não são pertinentes ao momento ou ao contexto do projeto. Essa etapa de curadoria é parte integrante do fluxo de trabalho com LLMs e não deve ser tratada como opcional. No experimento, mesmo com essa verificação e refinamento incluídos na tarefa, o tempo total permaneceu 25,8% menor na condição com LLM, sugerindo saldo temporal positivo no contexto estudado. Contudo, duas ressalvas se impõem. Primeiro, o tempo de verificação não foi medido separadamente do tempo de produção; trabalho futuro deve instrumentar essa decomposição. Segundo, é plausível que uma verificação mais rigorosa, capaz de elevar a corretude acima dos 84,1% observados, consumisse tempo adicional e reduzisse o ganho líquido. Em contextos críticos, nos quais a corretude requerida se aproxima de 100%, o custo de verificação pode ser substancialmente maior, e o benefício temporal do LLM, menor. Isso reforça a diretriz de que a validação humana

criteriosa é componente indispensável, e não opcional, do fluxo de trabalho com LLMs.

6.3 Ameaças À Validade

Validade interna. O efeito de aprendizagem entre rodadas foi mitigado pelo contrabalanceamento e pelo uso de cenários em domínios distintos com estrutura idêntica; o efeito de fadiga, pela curta duração de cada rodada (12 min); e a contaminação visual entre participantes, pelo uso de identificadores individuais e fiscalização dos aplicadores. Permanece risco de viés de seleção: participantes eram voluntários de comunidades de tecnologia, possivelmente mais entusiastas de IA, o que pode inflar desempenho e utilidade percebida. O risco de adivinhação de hipótese (*hypothesis guessing*) também não pode ser descartado: participantes podem ter inferido o objetivo do estudo e ajustado seu comportamento na condição com LLM.

Validade externa. O experimento foi realizado em evento técnico regional, com tarefas curtas (12 min) e cenários de ∼350 palavras; sessões reais são mais longas e envolvem interação direta com *stakeholders*. A generalização requer cautela e replicações. A diversidade da amostra (experiência de 1 a 26 anos) atenua, mas não elimina, essa limitação. O registro da versão/data do ChatGPT mitiga, mas não elimina, a validade temporal frente à rápida evolução dos LLMs.

Validade de construto. Há risco de mono-operação: utilizou-se um único LLM (ChatGPT), um único *prompt* e dois cenários, o que restringe o construto "uso de LLM" às escolhas operacionais feitas. A qualidade da interpretação foi operacionalizada por contagem contra *gold standard*; outras dimensões, como profundidade da justificativa, não foram capturadas. Esses riscos foram parcialmente mitigados pela validação dos cenários e da rubrica por especialistas, pela dupla avaliação cega e pela concordância aferida ($\kappa \geq 0{,}70$). O TAM aplicado apenas após a condição com LLM impede comparação de aceitação entre condições.

Validade de conclusão. O poder estatístico foi assegurado a priori (G*Power; n mínimo = 34; n real = 35). A violação de pressupostos foi tratada por verificação de normalidade e testes não-paramétricos quando apropriado. O risco de achados espúrios por múltiplas comparações (*fishing*) foi endereçado pela verificação de robustez à correção de Bonferroni (Seção 5.3), à qual todos os resultados resistiram.

7 Trabalhos Relacionados

A literatura recente sobre LLMs em ER pode ser organizada em três frentes: (i) avaliação de LLMs como agentes autônomos em tarefas de ER; (ii) uso de LLMs em subtarefas específicas (classificação, geração, detecção); e (iii) revisões e mapeamentos sistemáticos.

Na primeira frente, Ronanki et al. [4] realizaram o estudo mais diretamente relacionado, comparando saídas autônomas do ChatGPT com especialistas humanos em atributos de qualidade de requisitos. Encontraram superioridade do LLM em consistência e atomicidade, mas deficiência em não-ambiguidade, atributo em que o presente estudo observou resultado oposto quando o ChatGPT atua como ferramenta de apoio com *prompt* estruturado. Görer e Aydemir [13] exploraram o uso do ChatGPT para geração de roteiros de entrevista de elicitação a partir de descrições em linguagem natural, reportando resultados promissores mas destacando dependência da formulação do *prompt* e necessidade de curadoria humana. Luitel et al. [14] avaliaram o uso de LLMs para melhoria da completude de especificações de requisitos, encontrando aumento na cobertura de aspectos negligenciados; a limitação principal identificada foi a introdução de itens não fundamentados no texto original, observação convergente com o *trade-off* entre geração e verificação discutido na Seção 6.2.

Na segunda frente, Ferrari et al. [8] investigaram o uso de NLP para análise de requisitos em linguagem natural, identificando limitações no tratamento de ambiguidades contextuais que motivaram, em parte, a hipótese central deste estudo. Trabalhos subsequentes exploraram LLMs para classificação de requisitos funcionais versus não funcionais e detecção de conflitos, geralmente relatando desempenho competitivo com abordagens supervisionadas tradicionais, mas restritos a avaliações *offline* sobre *benchmarks*.

Na terceira frente, Arora et al. [3] revisam sistematicamente o uso de IA generativa em ER, identificando aplicações em geração de histórias de usuário, classificação de requisitos e detecção de conflitos, e destacando o papel moderador da qualidade dos *prompts*, observação que fundamentou o *prompt* padronizado adotado neste estudo.

Em conjunto, os estudos citados avaliam predominantemente o desempenho autônomo do LLM ou seu uso como gerador em subtarefas específicas, com evidência empírica ainda limitada sobre o efeito do LLM como ferramenta de apoio ao raciocínio humano em atividades de interpretação de requisitos sob condições experimentais controladas com desenho pareado. É essa lacuna que o presente artigo busca preencher, contribuindo com evidência quantitativa sobre o efeito conjunto do LLM em métricas comparativas (completude, corretude, identificação de problemas, tempo, carga cognitiva) e sobre a percepção de utilidade (TAM) em contexto ecológico.

8 Conclusão

Este artigo relatou um experimento controlado intra-sujeitos com 35 participantes, planejado e reportado segundo diretrizes consolidadas de experimentação em Engenharia de Software [5, 9], para avaliar o efeito do uso do ChatGPT na interpretação de requisitos durante atividades de elicitação. Os resultados fornecem evidência de que, no contexto estudado, o uso do LLM como apoio esteve associado a melhorias significativas nas cinco métricas comparativas avaliadas: completude (+11,9 p.p.), corretude (+8,5 p.p.), identificação de problemas (+1,3 itens), tempo (−25,8%) e carga cognitiva (−13,2 pontos), com tamanhos de efeito grandes e robustez à correção de Bonferroni. A utilidade percebida do uso do LLM foi elevada (M = 5,96/7,0).

Os achados sugerem, sem generalizar além do contexto experimental, que *prompts* estruturados que solicitam explicitamente a categorização de ambiguidades e omissões podem potencializar a atenção do engenheiro a esses elementos, invertendo, no modo de apoio, a limitação relatada na literatura para LLMs autônomos. Ao mesmo tempo, a corretude inferior a 100% e o compromisso entre geração e verificação (Seção 6.2) reforçam que a validação humana criteriosa permanece indispensável: o LLM potencializa o profissional, mas não o substitui.

Como trabalho futuro, pretende-se: replicar o experimento com outros LLMs e formatos de *prompt*; instrumentar a decomposição do tempo em produção e verificação; investigar de forma confirmatória o efeito moderador da experiência em ER; e conduzir estudos em contextos industriais com sessões de elicitação reais.

Disponibilidade de Artefatos

Os artefatos do estudo (transcrições dos cenários, gold standard, prompt padronizado, rubrica de avaliação, questionários NASA-TLX/TAM e dados anonimizados) serão disponibilizados publicamente após a publicação do artigo.

Agradecimentos

Os autores agradecem aos 35 participantes do estudo pela disponibilidade e engajamento durante o experimento.

Declaração de uso de IA: a IA generativa Claude (Anthropic) apoiou a redação, estruturação e formatação do manuscrito, incluindo parágrafos, Tabela 1 e Figura 1, a partir do protocolo e dados definidos pelos autores, que validaram todo o conteúdo e assumem responsabilidade pelo artigo.

Referências

[1] Boehm, B. (2006). A View of 20th and 21st Century Software Engineering. *Proc. ICSE*, 12–29.
[2] Vazquez, C. E.; Simões, G. S. (2016). *Engenharia de Requisitos*. Rio de Janeiro: Brasport.
[3] Arora, C.; Grundy, J.; Abdelrazek, M. (2024). Advancing Requirements Engineering via Generative AI. In: *Generative AI for Effective Software Development*. Springer, 129–148.
[4] Ronanki, K.; Berger, C.; Horkoff, J. (2023). Investigating ChatGPT's Potential to Assist in Requirements Elicitation Processes. arXiv:2307.07381.
[5] Wohlin, C. et al. (2012). *Experimentation in Software Engineering*. Berlin: Springer.
[6] Hart, S. G.; Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index). *Advances in Psychology*, 52, 139–183.
[7] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
[8] Ferrari, A. et al. (2017). Natural Language Requirements Processing: A Work Roadmap. *Proc. ICSSP*, 69–75.
[9] Jedlitschka, A.; Pfleeger, S. L. (2008). Reporting Experiments in Software Engineering. In: *Guide to Advanced Empirical Software Engineering*. Springer, 201–228.
[10] Montgomery, D. C. (2017). *Design and Analysis of Experiments*. 9th ed. Hoboken: Wiley.
[11] Shadish, W. R.; Cook, T. D.; Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
[12] Field, A.; Hole, G. (2003). *How to Design and Report Experiments*. London: Sage.
[13] Görer, B.; Aydemir, F. B. (2023). Generating Requirements Elicitation Interview Scripts with Large Language Models. *Proc. IEEE RE Workshops*, 44–51.
[14] Luitel, D.; Hassani, S.; Sabetzadeh, M. (2023). Using Language Models for Enhancing the Completeness of Natural-language Requirements. *Proc. REFSQ*, 87–104.