Avaliação da integração do método colaborativo do USARP com LLM

Andreza H. Magalhães, Ana Emilly A. Oliveira, Valéria M. Pinheiro e Anna Beatriz S. Marques

[1]Universidade Federal do Ceará (UFC) - Campus Russas
Russas - CE - Brasil

andrezahm@alu.ufc.br, ana.emilly.147@gmail.com

{valeria.pinheiro, beatriz.marques}@ufc.br

Abstract. *Collaborative software requirements elicitation is essential to system quality but faces cognitive and coordination challenges among participants. Structured methods, such as USARP, aim to mitigate these difficulties through collaborative practices. This paper investigates the impact of integrating Large Language Models (LLMs) into USARP, considering different prompt engineering strategies. A controlled experiment compared the traditional application of the method with the use of LLMs under Zero-Context and Context-Rich approaches. The results indicate improvements in requirements quality, particularly with enriched prompts, corroborating studies on the influence of context on LLM performance.*

Resumo. *A elicitação colaborativa de requisitos de software é essencial para a qualidade dos sistemas, mas enfrenta desafios cognitivos e de coordenação entre participantes. Métodos estruturados, como o USARP, buscam mitigar essas dificuldades por meio de práticas colaborativas. Neste trabalho, investigamos o impacto da integração de Modelos de Linguagem de Grande Escala (LLMs) ao USARP, considerando diferentes estratégias de engenharia de prompts. Um experimento controlado comparou a aplicação tradicional do método com o uso de LLMs em abordagens Zero-Context e Context-Rich. Os resultados indicam melhoria na qualidade dos requisitos, especialmente com prompts enriquecidos, corroborando estudos sobre a influência do contexto no desempenho de LLMs.*

1. Introdução

Modelos de Linguagem de Grande Escala (Large Language Models – LLMs) têm sido explorados na Engenharia de Requisitos como apoio à elicitação e ao refinamento de requisitos, com resultados promissores em termos de eficiência e qualidade dos artefatos produzidos [Kasauli et al. 2021, Hou et al. 2024]. No contexto da elicitação de requisitos de usabilidade, métodos colaborativos estruturados, como o USARP (*Usability Requirements with Personas and user stories*), têm se mostrado eficazes ao integrar diferentes perspectivas por meio do uso de personas, histórias de usuário e sessões colaborativas de brainstorming [de Oliveira et al. 2020, Marques et al. 2023]. Nesse cenário, os LLMs emergem como potenciais mediadores sociotécnicos, sendo um elemento que faz a ponte entre as pessoas (dimensão social) e os artefatos, regras e tecnologias (dimensão técnica), influenciando como a interação acontece e o que é produzido a partir dela [Valadão et al.

2014, Abbasi et al. 2025], capazes de apoiar processos colaborativos e atividades cognitivamente complexas, especialmente em grupos com níveis variados de experiência [Hemmat et al. 2025].

Apesar de seus benefícios, o método USARP pode demandar esforço significativo e conhecimento especializado em usabilidade, o que pode impactar a consistência e a completude dos requisitos gerados, sobretudo em contextos educacionais ou com participantes menos experientes [Fonseca et al. 2024]. Embora os LLMs apresentem potencial para mitigar esses desafios, ainda há poucas evidências empíricas sobre os efeitos de sua integração com métodos colaborativos estruturados na elicitação de requisitos de usabilidade. Além disso, estudos indicam que o desempenho desses modelos é fortemente influenciado pelas estratégias de engenharia de *prompts* adotadas [Brown et al. 2020, Huang et al. 2025]. Em particular, o impacto de diferentes níveis de contexto fornecidos aos LLMs, como *prompts* com pouco ou muito contexto, permanece pouco explorado nesse domínio.

Com o intuito de investigar o impacto da integração de LLMs ao método USARP na elicitação de requisitos de usabilidade, temos como objetivos: (i) avaliar a qualidade dos requisitos gerados com o apoio de LLMs em comparação ao uso exclusivo do USARP; (ii) comparar estratégias de *prompting Zero-context* e *Context-Rich*; e (iii) analisar o papel dos LLMs como mediadores sociotécnicos em processos colaborativos de elicitação. Para atingií-los, foi conduzido um experimento controlado considerando três condições: USARP sem o uso de LLM, USARP com LLM utilizando *prompts Zero-context* e USARP com LLM utilizando *prompts Context-Rich*.

Este artigo está organizado em 8 seções, sendo a Seção 1 a introdução, que expõe o contexto, a problemática, o objetivo do estudo e a solução proposta. A Seção 2 discute os trabalhos relacionados ao tema. Na Seção 3 é apresentada a fundamentação teórica, abordando a metodologia USARP e o processo de elicitação de requisitos a partir dessa abordagem. Na Seção 4 propusemos uma solução para a integração apresentada. A metodologia adotada, incluindo a condução do experimento e a comparação entre a elicitação sem LLM e com LLM, bem como entre os tipos de *prompts* utilizados, foi abordada na Seção 5. Os resultados obtidos são apresentados e analisados na Seção 6, destacando os testes realizados, a comparação entre os métodos avaliados e uma breve discussão sobre as possíveis limitações desse estudo. A Seção 7 apresenta a conclusão do artigo, com base nos resultados alcançados. E, por conseguinte, nossos agradecimentos são apresentados na Seção 8.

2. Trabalhos Relacionados

A pesquisa em Sistemas Colaborativos evidencia que a elicitação de requisitos vai além de aspectos técnicos, envolvendo dimensões sociais e cognitivas. Estudos como os de Stefani e Duduchi (2023) e Diniz et al. (2025) destacam a importância de habilidades colaborativas, enquanto Mantau e Benitti (2023) e Gonçalves et al. (2024) reforçam a necessidade de ferramentas que apoiem a comunicação, a coordenação e a construção compartilhada de conhecimento.

No contexto da Engenharia de Requisitos com LLMs, Dalpiaz e Niu (2020) apontam o potencial da IA no apoio à análise e à qualidade dos requisitos, enquanto Vogelsang (2024) destaca o papel central do *prompting*. Trabalhos recentes indicam que LLMs

podem apoiar a geração, análise e refinamento de *user stories*, com resultados dependentes da qualidade dos *prompts*, embora ainda apresentem limitações [Quattrocchi et al. 2026, Garcia et al.].

Nesse cenário, este artigo contribui ao investigar empiricamente a integração de LLMs ao método colaborativo USARP, analisando seu papel como mediadores sociotécnicos e o impacto de diferentes estratégias de *prompting* na qualidade dos requisitos e na dinâmica colaborativa.

3. Fundamentação Teórica

3.1. Elicitação Colaborativa de Requisitos de Software

A elicitação colaborativa de requisitos de software envolve a coordenação de múltiplos participantes com diferentes perfis de conhecimento, visando capturar necessidades de usuários e partes interessadas de forma compartilhada. No contexto de métodos ágeis e coletivas, Stefani e Duduchi (2023) , destacam que elementos colaborativos como: comunicação, coordenação e cooperação, são parte essencial para que as equipes alinhem expectativas, negociem soluções e construam requisitos de forma integrada. Além disso, o desenvolvimento compartilhado de software exige não apenas competências técnicas, mas também competências sociais e cognitivas, como comunicação clara e tomada de decisão conjunta, que influenciam diretamente a eficácia da elicitação de requisitos [Diniz et al. 2025]. Esses estudos evidenciam que a heterogeneidade de habilidades e perspectivas dos participantes pode tanto enriquecer o processo quanto gerar desafios no estabelecimento de consenso, na gestão da carga cognitiva e na tradução de percepções subjetivas em requisitos bem definidos, reforçando a necessidade de métodos e ferramentas que suportem a colaboração de maneira estruturada.

3.2. O método USARP

O método USARP é um método colaborativo voltado à elicitação de requisitos de usabilidade que integra personas, histórias de usuário e sessões estruturadas de brainstorming com equipes heterogêneas. Seu objetivo é apoiar a identificação de requisitos de usabilidade, promovendo a participação ativa dos envolvidos e a construção compartilhada dos artefatos [de Oliveira et al. 2020].

Estudos empíricos indicam que o USARP pode ser aplicado tanto em contextos educacionais quanto industriais, contribuindo para a melhoria da qualidade dos requisitos elicitados, especialmente em termos de clareza e alinhamento com as necessidades dos usuários [Marques et al. 2023]. O caráter colaborativo do método favorece a integração de diferentes perspectivas e reduz ambiguidades em processos de elicitação pouco estruturados, reforçando seu potencial como abordagem de apoio à colaboração em Engenharia de Requisitos.

Segundo Silva et al. (2024) , o método USARP é estruturado em cinco etapas principais voltadas ao refinamento coletivo de *user stories* e de requisitos de usabilidade, conforme ilustrado na Figura 1. Essas etapas organizam o processo de elicitação e refinamento, promovendo a participação ativa dos envolvidos e o uso de artefatos de apoio.

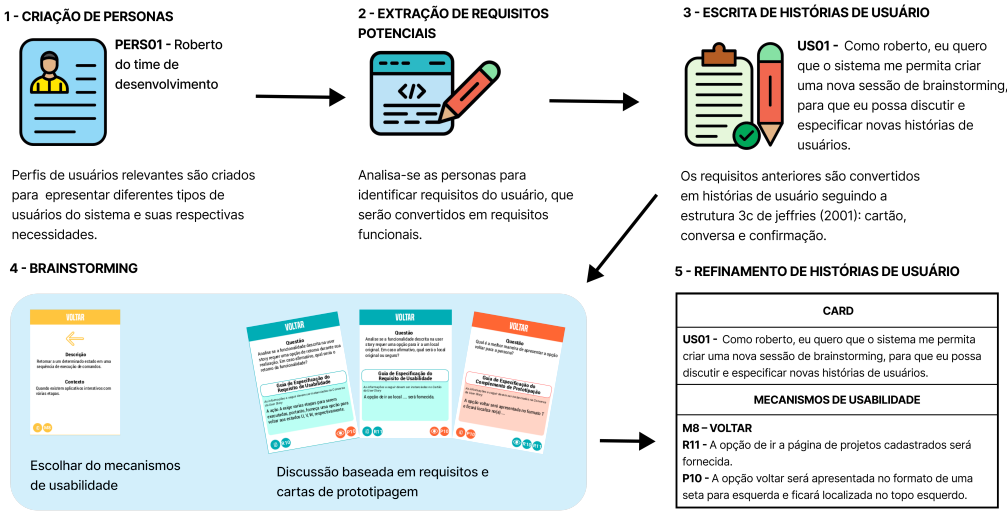


Figura 1. Fluxo de atividades do método USARP [Silva et al. 2024].

1. **Desenvolvimento de personas:** para orientar soluções de usabilidade, o USARP recomenda a criação de personas que representem os usuários, considerando suas necessidades e características [Silva et al. 2024];
2. **Identificação de requisitos potenciais:** a equipe deve identificar requisitos que atendam às necessidades dos usuários, analisando informações das personas e outros dados coletados durante a pesquisa com usuários [Silva et al. 2024];
3. **Definição de *user stories*:** a partir dos requisitos identificados, a equipe especifica um conjunto de *user stories*. O USARP adota o modelo 3C (*Card, Conversation, Confirmation*) para estruturar essas histórias, sendo o foco inicial a elaboração do *Card* [Silva et al. 2024];
4. **Sessões de *brainstorming*:** a equipe explora mecanismos de usabilidade por meio de *user stories* e sessões de *brainstorming*, usando as cartas do USARP como guia e envolvendo profissionais de diferentes áreas para garantir múltiplas perspectivas [Silva et al. 2024];
5. **Refinamento das *user stories*:** as decisões do *brainstorming* são registradas na seção de ”Mecanismos de Usabilidade”de cada *user story* [Silva et al. 2024].

3.3. LLMs e Engenharia de Prompts

Os LLMs vêm sendo amplamente investigados como ferramentas de apoio a atividades cognitivamente complexas na Engenharia de Software, incluindo tarefas que envolvem análise, síntese e tomada de decisão em contextos colaborativos. Diversos estudos indicam que o desempenho desses modelos depende fortemente da forma como as instruções são formuladas, o que torna a engenharia de *prompts* um elemento central para a obtenção de respostas adequadas [Brown et al. 2020, Liu et al. 2023]. Evidências apresentadas por Wei et al. (2022) e Zamfirescu-Pereira et al. (2023) apontam que *prompts* enriquecidos com maior contexto tendem a produzir respostas mais consistentes, coerentes e alinhadas a objetivos específicos, especialmente em tarefas que demandam interpretação estruturada e aderência a diretrizes, como a elicitação e o refinamento de requisitos de software. Esses resultados reforçam a relevância do desenho cuidadoso de *prompts* na integração de LLMs a processos colaborativos, nos quais a informação gerada impacta diretamente a construção coletiva de artefatos.

4. Solução Proposta

Com essa proposta de integrar LLMs ao método USARP para apoiar a elicitação colaborativa de requisitos de usabilidade, os LLMs atuam como mediadores sociotécnicos, auxiliando os participantes na geração e no refinamento dos requisitos a partir de insumos estruturados do método. A integração é realizada por meio da engenharia de *prompts*, que traduz elementos do USARP, como personas e histórias de usuário, em instruções fornecidas ao LLM, conforme ilustrado na Figura 2.

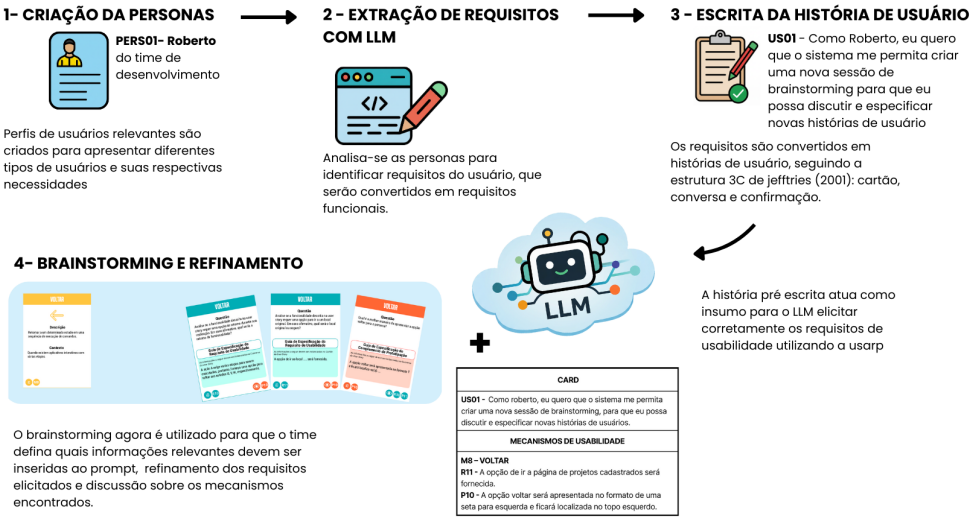


Figura 2. Fluxo de atividades do método USARP com o auxílio de LLMs

Foram investigadas duas estratégias de *prompting*: *Zero-context*, com insumos mínimos, e *Context-Rich*, com informações adicionais e orientações estruturadas. Dessa forma, a solução busca reduzir o esforço dos participantes durante a elicitação, apoiar grupos com diferentes níveis de experiência e fortalecer a colaboração da IA com a USARP, na elicitação de requisitos.

5. Metodologia

Este estudo adotou uma abordagem **quantitativa e experimental**, conduzida por meio de um experimento controlado, conforme as diretrizes propostas por Wohlin et al. (2012), com o objetivo de investigar o impacto da integração do ChatGPT (GPT-5) ao método USARP, bem como a influência de diferentes estratégias de engenharia de *prompts*.

5.1. Design do Experimento

O estudo foi planejado e executado como um experimento controlado aplicado na disciplina de Introdução a Processos e Requisitos de Software (IPRS), ofertada para os cursos de Engenharia de Software e Ciência da Computação, a partir do segundo semestre da graduação, na Universidade Federal do Ceará, Campus Russas. O experimento foi conduzido ao longo de dois dias, reunindo estudantes regularmente matriculados na disciplina. No primeiro encontro, realizado no dia 23 de junho de 2025, participaram 46 estudantes, e no segundo encontro, ocorrido em 26 de junho de 2025, contamos com a participação de 50 estudantes.

A escolha pelo delineamento em dois momentos teve como objetivo criar condições de comparação entre diferentes formas de elicitação de requisitos de usabilidade, sendo elas a elicitação com o método do USARP, mas sem o auxílio de LLMs e a elicitação com o auxílio dessa ferramenta. Além disso, ao longo das duas etapas, foram coletados artefatos específicos com o propósito de viabilizar análises posteriores, incluindo o formulário de Perfil do Participante, o formulário de *Feedback* do Experimento e os *prompts* utilizados durante a elicitação. Esses registros serviram como insumos para análises quantitativas, realizadas por meio de métodos estatísticos.

Hipóteses do experimento

Com base no desenho experimental adotado, foram formuladas as seguintes hipóteses para orientar a análise estatística:

- **Hipótese A – Técnica de Elicitação:**
  - $\mathbf{H}_{01}$: não há diferença significativa na elicitação de requisitos com o método USARP utilizando LLMs.
  - $\mathbf{H}_{11}$: há diferença significativa na elicitação de requisitos com o método USARP quando apoiado por LLMs.
- **Hipótese B – Tipo de *Prompt*:**
  - $\mathbf{H}_{02}$: não há diferença significativa na qualidade das respostas geradas pela LLM quando o usuário fornece insumos mínimos (*Zero-context prompting)* ou insumos completos (*Context Rich prompting*).
  - $\mathbf{H}_{12}$: há diferença significativa na qualidade das respostas quando insumos completos são fornecidos à LLM, aumentando a completude e consistência dos requisitos.

Essas hipóteses refletem diretamente os dois fatores do experimento fatorial 2x2, considerando a técnica de elicitação (com e sem LLM) e o tipo de *prompting (Zero-context vs. Context Rich)*.

Em cada experimento foram disponibilizados formulários para nos auxiliarem na coleta de dados para as nossas análises, sendo eles: Perfil do Participante, *Feedback* do Experimento, que acompanhavam o Termo de Consentimento Livre e Esclarecido (TCLE) em ambos os formulários, e *Prompts* do Experimento, que auxiliaram na realização das nossas análises quantitativas.

A pesquisa foi conduzida em conformidade com os princípios éticos aplicáveis a estudos com seres humanos, tendo sido aprovada pelo Comitê de Ética em Pesquisa (CEP), sob o CAAE nº 86282124.0.0000.5054 e, seguindo as diretrizes éticas com a participação voluntária dos estudantes, conforme esclareido no TCLE, e todos os dados sensíveis foram anonimizados para a análise.

5.2. Primeiro dia do experimento (sem LLM)

O encontro teve início com uma breve apresentação dos objetivos do estudo e do método USARP. Em seguida, os estudantes foram organizados em nove grupos heterogêneos, com pelo menos dois participantes, e receberam materiais contendo um cenário fictício, uma persona associada a uma *User Story* (US), Regras de Negócio (RN) e o *checklist* do método, além de acesso às cartas da USARP.

A atividade consistiu na construção coletiva de requisitos de usabilidade, a partir desses insumos, com base na análise da relação entre os elementos fornecidos. Os requisitos produzidos, sem o apoio de LLMs, constituíram a base de comparação para a segunda etapa do experimento.

A segunda etapa foi realizada em laboratório de informática, com o uso da LLM como ferramenta central no processo de elicitação de requisitos. Inicialmente, os participantes receberam uma breve introdução à engenharia de *prompts*, com explicação das abordagens *Zero-Context prompting*, baseada em insumos limitados, e *Context Rich prompting*, caracterizada pelo uso de múltiplos insumos.

Em seguida, os estudantes foram divididos em dois grupos experimentais, cada um associado a uma dessas estratégias, e organizados em subgrupos de até três integrantes para favorecer a colaboração na elaboração dos *prompts*. Foi utilizada exclusivamente a ferramenta ChatGPT (GPT-5), a fim de padronizar as condições do experimento.

- **Grupos *Zero-Context prompting***: receberam um cenário fictício impresso, contendo a persona, a US e as RNs do sistema, além do PDF das cartas do método USARP. Orientamos a esses grupos que repassassem à LLM como insumo, apenas o PDF das cartas e elaborassem um *prompt* que fizesse com que a LLM os entregassem as cartas necessárias para a US oferecida.
- **Grupos *Context Rich prompting***: além do cenário fictício e do PDF das cartas, esses grupos também receberam o PDF do checklist do método USARP e o documento de descrição do método. A orientação era criar um *prompt*, que incluísse instruções explícitas sobre a tarefa e disponibilizasse à LLM todos os insumos fornecidos.

Observação: Os cenários dos dias 1 e 2 eram diferentes, porém possuiam a mesma padronagem.

Cada grupo foi responsável por elaborar seu próprio *prompt*, refletindo em conjunto sobre quais informações eram mais relevantes e como apresentá-las à LLM. Durante a execução, os participantes se envolveram em discussões acerca da clareza dos *prompts*, da suficiência das informações repassadas e da forma como a LLM estruturava as respostas.

Ao final da atividade, foram recolhidos os artefatos planejados, incluindo os *links* das conversas realizadas com a LLM, os formulários de *Feedback* do Experimento, nos quais os estudantes relataram suas percepções sobre a experiência, e os formulários de *Prompts* do Experimento, contendo os registros dos *prompts* elaborados.

5.3. Análise de dados

O experimento teve como objetivo comparar os resultados obtidos no primeiro e no segundo dia de execução, bem como analisar o formato dos *prompts* gerados na segunda etapa. Inicialmente, foram identificados 36 estudantes presentes em ambos os dias; no entanto, foram observados casos em que participantes atuaram individualmente durante a execução das atividades. Considerando que o delineamento do estudo previa a realização das tarefas em grupo e que o método USARP possui natureza colaborativa, optou-se pela exclusão desses casos, uma vez que a atuação individual não representa adequadamente

as dinâmicas de interação e construção coletiva esperadas. Dessa forma, a amostra final foi composta por 30 participantes.

A análise concentrou-se na comparação dos requisitos produzidos com e sem o uso de LLMs, considerando os critérios de corretude, completude, consistência e adequação ao método USARP. Os resultados dos grupos foram atribuídos individualmente a cada participante e, posteriormente, comparados entre o primeiro e o segundo dia em nível individual, desconsiderando a composição dos grupos, que foi alterada entre as etapas.

A avaliação envolveu dados quantitativos, a partir de dois documentos de requisitos: um elaborado sem LLMs e outro com apoio de LLMs. Para isso, utilizou-se um formulário estruturado com quatro categorias avaliativas e seis questões validadoras, disponível nos Materiais do experimento – SBSC 2026. Os resultados foram classificados em escala ordinal de 0 a 5, em que 0 indica impossibilidade de comparação e 5 indica total atendimento ao critério avaliado.

As avaliações foram realizadas por duas analistas de requisitos com aproximadamente 18 meses de experiência no método USARP, adotando procedimento de consenso em caso de divergências. O protocolo permitiu mensurar sistematicamente a qualidade dos requisitos e analisar o impacto do uso de LLMs, considerando diferentes níveis de contexto fornecido aos *prompts*.

6. Resultados

Esta seção apresenta os resultados do experimento controlado que avaliou o impacto do uso de LLMs e do nível de contexto nos *prompts* durante a elicitação de requisitos, considerando as dimensões de Corretude, Completude, Consistência e Adequação. As análises combinam estatísticas descritivas, testes não paramétricos para amostras pareadas e independentes, e visualizações por meio de *raincloud plots*.

6.1. Impacto do Uso da LLM

Inicialmente, foi realizada uma análise descritiva dos dados para caracterizar o desempenho dos participantes em cada dimensão avaliada. As estatísticas descritivas são apresentadas na Figura 3a, que detalha medidas de tendência central e dispersão para os cenários com LLM e sem LLM, bem como com contexto e sem contexto.

	N	Mean	SD	SE	Coefficient of variation
CorretudeSL	30	1.867	1.279	0.234	0.685
CorretudeCL	30	3.167	1.315	0.240	0.415
CompletudeSL	30	1.333	1.295	0.237	0.972
CompletudeCL	30	3.300	1.579	0.288	0.478
ConsistênciaSL	30	1.233	1.278	0.233	1.036
ConsistênciaCL	30	3.900	1.296	0.237	0.332
AdequaçãoSL	30	1.900	1.062	0.194	0.559
AdequaçãoCL	30	3.467	1.634	0.298	0.471

(a) Indicadores descritivos sem LLM vs. com LLM

			W	p
CorretudeSL	-	CorretudeCL	0.926	.039
CompletudeSL	-	CompletudeCL	0.956	.243
ConsistênciaSL	-	ConsistênciaCL	0.880	.003
AdequaçãoSL	-	AdequaçãoCL	0.887	.004

Nota. Significant results suggest a deviation from normality.

(b) Teste Shapiro-Wilk

Figura 3. Resultados da análise descritiva e e teste de normalidade - Com e Sem LLM

Observou-se que o grupo CL (com LLM) apresentou médias superiores ao SL (sem LLM) em todas as dimensões avaliadas (Corretude, Completude, Consistência e

Adequação). As diferenças foram mais pronunciadas em Consistência e Completude. Ademais, o grupo SL apresentou maior coeficiente de variação, indicando maior dispersão dos dados em comparação ao CL, que demonstrou desempenho mais elevado e relativamente mais estável entre os participantes.

A avaliação da normalidade por meio do teste de Shapiro–Wilk (Figura 3b) indicou rejeição da hipótese nula de distribuição normal para as variáveis Corretude ($p = 0{,}039$), Consistência ($p = 0{,}003$) e Adequação ($p = 0{,}004$), evidenciando desvio significativo da normalidade. Para Completude, não houve evidência estatística suficiente para rejeitar a normalidade ($p = 0{,}243$). Considerando a violação do pressuposto de normalidade em parte das variáveis, adotou-se procedimento não paramétrico nas análises inferenciais subsequentes.

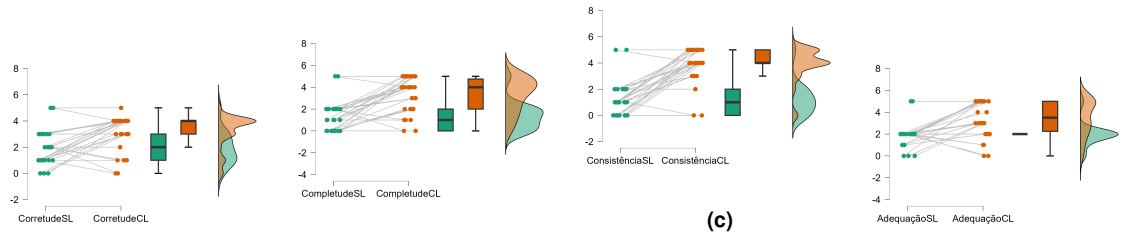
O teste de Wilcoxon indicou diferenças estatisticamente significativas entre os dados SL e CL, com $p < 0{,}001$. Os valores de z variaram entre $-4{,}52$ e $-4{,}76$, evidenciando um efeito consistente na mesma direção. As correlações bisseriais por postos (variando aproximadamente de $-0{,}79$ a $-0{,}86$) indicam magnitude de efeito elevada, sugerindo superioridade sistemática do grupo CL em relação ao SL (Figura 4).

									95% CI for Rank-Biserial Correlation	
Measure 1		Measure 2	W	z	df	p	Rank-Biserial Correlation	SE Rank-Biserial Correlation	Lower	Upper
CorretudeSL	-	CorretudeCL	33.000	−3.484		< .001	−0.797	0.225	−0.912	−0.566
CompletudeSL	-	CompletudeCL	16.500	−4.038		< .001	−0.906	0.221	−0.960	−0.788
ConsistênciaSL	-	ConsistênciaCL	4.000	−4.445		< .001	−0.979	0.217	−0.991	−0.951
AdequaçãoSL	-	AdequaçãoCL	27.000	−3.646		< .001	−0.834	0.225	−0.929	−0.637

Nota. Wilcoxon signed-rank test.

Figura 4. Teste de Wilcoxon

Os *raincloud plots* evidenciam deslocamento consistente das distribuições do grupo CL em relação ao SL em todas as dimensões avaliadas. Observa-se maior concentração de valores em faixas superiores para o CL, com menor sobreposição entre as distribuições, especialmente em Completude (Figura 5b) e Consistência (Figura 5c). Os gráficos também indicam menor dispersão relativa no CL e maior variabilidade no SL, corroborando os resultados descritivos e inferenciais previamente apresentados Figura 5.



(a) Corretude **(b) Completude** **(c) Consistência** **(d) Adequação**

Figura 5. Raincloud Plots - Análises Com e Sem LLM

6.2. Impacto do Contexto na LLM

A análise descritiva dos dados revela uma superioridade sistemática do grupo *Context Rich* em todas as métricas avaliadas, com destaque para a variável Consistência, que atin-

giu média de $4{,}500$ frente aos $3{,}214$ do grupo *Zero-Context*. A dispersão dos dados, aferida pelo coeficiente de variação, foi superior no grupo sem contexto, alcançando $0{,}570$ na métrica de Adequação, indicando desempenho mais instável e heterogêneo na ausência de informações contextuais. Essa tendência de maior precisão no grupo contextualizado é corroborada pelos postos médios (*Mean Rank*), que se mantêm consistentemente elevados em Corretude $(18{,}41)$ e Consistência $(19{,}50)$ (Figura 6a).

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation	Mean Rank	Sum Rank
Corretude	Context Rich	16	3.563	1.094	0.273	0.307	18.41	294.5
	Zero-Context	14	2.714	1.437	0.384	0.530	12.18	170.5
Completude	Context Rich	16	3.438	1.548	0.387	0.450	16.41	262.5
	Zero-Context	14	3.143	1.657	0.443	0.527	14.46	202.5
Consistência	Context Rich	16	4.500	0.516	0.129	0.115	19.50	312.0
	Zero-Context	14	3.214	1.578	0.422	0.491	10.93	153.0
Adequação	Context Rich	16	3.750	1.483	0.371	0.396	16.88	270.0
	Zero-Context	14	3.143	1.791	0.479	0.570	13.93	195.0

(a) Indicadores descritivos *Zero-context vs. Context Rich*

Residuals	W	p
Corretude	0.788	< .001
Completude	0.885	.004
Consistência	0.833	< .001
Adequação	0.891	.005

Nota. Significant results suggest a deviation from normality.

(b) Teste Shapiro-Wilk

Figura 6. Resultados da análise descritiva e do teste de normalidade - *Zero-context e Context Rich*

A fundamentação para o emprego de estatística não paramétrica decorre da violação do pressuposto de normalidade em todas as variáveis dependentes. O teste de Shapiro–Wilk (Figura 6b) aplicado aos resíduos resultou em valores de $p < 0{,}001$ para Corretude e Consistência, e $p < 0{,}005$ para Completude e Adequação, implicando a rejeição da hipótese nula de distribuição normal. Os *Raincloud Plots* ratificam visualmente essa condição, ao evidenciarem distribuições de densidade de *Kernel* com assimetria acentuada e concentrações multimodais, características que inviabilizam o uso de testes paramétricos, como o teste t de Student, para esta amostra.

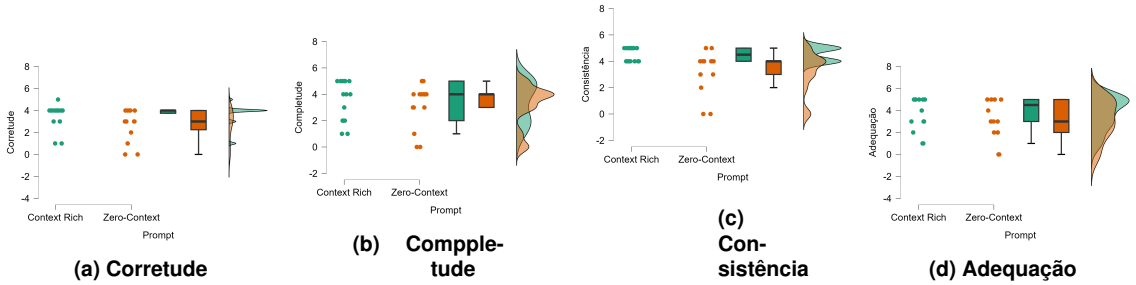
Independent Samples T-Test

						95% CI for Rank-Biserial Correlation	
	U	df	p	Rank-Biserial Correlation	SE Rank-Biserial Correlation	Lower	Upper
Corretude	158.5		.036	−0.415	0.212	−0.698	−0.021
Completude	126.5		.550	−0.129	0.212	−0.501	0.283
Consistência	176.0		.004	−0.571	0.212	−0.790	−0.225
Adequação	134.0		.347	−0.196	0.212	−0.551	0.218

Note. For the Mann-Whitney test, effect size is given by the rank biserial correlation.
Note. Mann-Whitney U test.

Figura 7. Teste Mann–Whitney U

O teste de Mann–Whitney U confirmou que o efeito do contexto é estatisticamente significativo e de magnitude relevante para Consistência ($p = 0{,}004$) e Corretude ($p = 0{,}036$). O tamanho do efeito, mensurado pela correlação bisserial por postos, atingiu $-0{,}571$ para Consistência, indicando associação robusta entre a presença de contexto e a estabilidade das respostas. Embora as variáveis Completude ($p = 0{,}550$) e Adequação ($p = 0{,}347$) tenham apresentado médias nominais superiores no grupo contextualizado, a análise inferencial demonstra que tais diferenças não alcançaram significância estatística, sugerindo que o contexto atua de forma mais decisiva na precisão e no rigor lógico do que na extensão ou adequação geral do conteúdo (Figura 7).



(a) Corretude (b) Compple-tude (c) Con-sistência (d) Adequação

Figura 8. Raincloud Plots - Análises Zero-Context e Context Rich LLM

A síntese visual proporcionada pelos Raincloud Plots (Figura 8) evidencia a redução drástica da variabilidade no grupo *Context Rich*, especialmente na dimensão de Consistência (Figura 8c), onde os dados se agrupam densamente no topo da escala. Em contrapartida, o grupo *Zero-Context* manifesta uma distribuição ”em chuva”(pontos individuais) muito mais esparsa, com a presença de valores baixos que deslocam a mediana para níveis inferiores. Essa configuração gráfica, somada aos intervalos de confiança da correlação bisserial de postos que não cruzam o zero para as variáveis significantes, consolida a conclusão de que o contexto funciona como um redutor de ruído e um indutor de qualidade técnica superior.

- Síntese dos Resultados

As análises estatísticas revelam que a utilização de modelos de linguagem (LLMs) com suporte de contexto (CL/*Context Rich*) supera significativamente o desempenho sem contexto (SL/*Zero-Context*) em múltiplas métricas de qualidade. Na comparação pareada (*Wilcoxon signed-rank test*), o grupo com contexto (CL) apresentou médias superiores e valores de $p < 0{,}001$ em todas as categorias — Corretude ($W = 33{,}0$), Completude ($W = 16{,}5$), Consistência ($W = 4{,}0$) e Adequação ($W = 27{,}0$) — com elevados tamanhos de efeito (correlação bisserial por postos variando de $-0{,}797$ a $-0{,}979$).

Na análise de amostras independentes (*Mann–Whitney U test*), a superioridade do contexto rico foi confirmada de forma estatisticamente significativa para Corretude ($p = 0{,}036$) e Consistência ($p = 0{,}004$), esta última apresentando a maior disparidade de médias ($4{,}50$ vs. $3{,}21$) e o menor coeficiente de variação ($0{,}115$), indicando maior estabilidade nos resultados sob condição contextualizada.

6.3. Ameaças à Validade

Como todo estudo experimental, esta pesquisa está sujeita a ameaças à validade, que foram analisadas com base nas categorias propostas por Wohlin et al. (2012) .

Validade interna. Refere-se à existência de fatores que possam ter influenciado os resultados além das variáveis investigadas. Como o experimento foi conduzido em ambiente controlado, buscou-se minimizar interferências externas por meio da padronização dos materiais, instruções e ferramentas utilizadas. Ainda assim, diferenças no nível de experiência dos participantes (sobre o método e sobre elicitação de requisitos) e na dinâmica dos grupos podem ter influenciado os resultados. Além disso, a avaliação dos requisitos, embora realizada por analistas com experiência e com procedimento de consenso, pode envolver certo grau de subjetividade.

Validade externa. Diz respeito à generalização dos resultados. O estudo foi conduzido com estudantes em contexto acadêmico, o que pode limitar a extrapolação dos achados para ambientes industriais. Embora esse público seja frequentemente utilizado em experimentos em Engenharia de Software, seus resultados podem não refletir integralmente a atuação de profissionais experientes. Ademais, o uso de uma única ferramenta (ChatGPT – GPT-5) pode restringir a generalização para outros modelos de LLMs.

Validade de constructo. Relaciona-se à adequação das métricas utilizadas para representar os conceitos investigados. A qualidade dos requisitos foi avaliada com base em critérios como corretude, completude, consistência e adequação, amplamente utilizados na literatura. Ainda assim, tais métricas podem não capturar completamente todos os aspectos da qualidade em contextos reais, especialmente aqueles relacionados à colaboração e à negociação entre participantes.

Validade de conclusão. Refere-se à confiabilidade das conclusões obtidas a partir dos dados. Para mitigar essa ameaça, foram utilizados testes estatísticos não paramétricos adequados ao comportamento dos dados, bem como múltiplas análises descritivas e inferenciais. No entanto, o tamanho da amostra e a variabilidade entre os participantes podem influenciar a robustez dos resultados.

Apesar dessas limitações, acredita-se que os cuidados metodológicos adotados contribuem para a confiabilidade dos achados e fornecem evidências relevantes sobre o uso de LLMs como apoio à elicitação colaborativa de requisitos.

7. Conclusões

Os resultados indicam que **a integração de LLMs ao método USARP contribui para a elicitação colaborativa de requisitos de usabilidade, resultando em artefatos mais consistentes quando comparados ao uso exclusivo do método**. A análise das estratégias de prompting mostrou que o fornecimento de maior contexto ao LLM impacta positivamente a qualidade dos requisitos, destacando a engenharia de *prompts* como elemento central nesse processo. Sob a perspectiva de Sistemas Colaborativos, os LLMs atuam como mediadores sociotécnicos, apoiando a colaboração e reduzindo barreiras cognitivas, especialmente em grupos com diferentes níveis de experiência. Logo, conclui-se que a integração de LLMs ao USARP é uma abordagem promissora ao explorar a IA como tecnologia para apoiar a colaboração, a inovação e a transformação de práticas na Engenharia de Software.

Como trabalhos futuros, sugere-se investigar o uso de LLMs em experiências práticas de ensino de Engenharia de Requisitos, avaliar diferentes modelos e suas estratégias de *prompting*, e analisar sua eficácia em contextos industriais reais envolvendo equipes multidisciplinares e *stakeholders* diversos. Essas linhas de pesquisa podem ampliar o entendimento sobre o papel da Inteligência Artificial na Engenharia de Requisitos, contribuindo para o avanço tanto da prática quanto da pesquisa em sistemas com enfoque na colaboração.

8. Agradecimentos

Este artigo contou com o apoio de ferramentas de Inteligência Artificial generativa (*ChatGPT* – OpenAI), utilizadas exclusivamente para revisão textual, organização de

seções, síntese de conteúdo e otimização na busca por referências. Os autores são integralmente responsáveis pelo conteúdo apresentado.

Referências

Abbasi, M. A., Ihantola, P., Mikkonen, T., and Mäkitalo, N. (2025). Towards human-ai synergy in requirements engineering: A framework and preliminary study. In *2025 Sixth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 81–88.

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.

Dalpiaz, F. and Niu, N. (2020). Requirements engineering in the days of artificial intelligence. *IEEE software*, 37(4):7–10.

de Oliveira, G. F., Ferreira, B., and Marques, A. B. (2020). Usarp method: eliciting and describing usability requirements with personas and user stories. In *Proceedings of the XXXIV Brazilian Symposium on Software Engineering*, pages 437–446.

Diniz, W., Gadelha, B., Steinmacher, I., and França, C. (2025). Habilidades colaborativas no mercado de ti: Uma investigação sobre requisitos de soft skills. In *Anais do XX Simpósio Brasileiro de Sistemas Colaborativos*, pages 139–150, Porto Alegre, RS, Brasil. SBC.

Fonseca, C., Zaina, L., and Marques, A. (2024). Avaliação de métodos para elicitação e especificação de requisitos de usabilidade com histórias de usuário: Um experimento controlado. In *Anais do XXXVIII Simpósio Brasileiro de Engenharia de Software*, pages 257–268, Porto Alegre, RS, Brasil. SBC.

Garcia, P., Depollo, J. V., Barroso, L., Figueiredo, E., Constantino, K., Fernandes, F., and Côgo, F. R. Documenting user stories: Does llm help?

Gonçalves, E., Lima, I., Sousa, J., Rocha, M., and Rabelo, J. (2024). Gestlab: Software de gestão do conhecimento e processos colaborativos no contexto universitário. In *Anais do XIX Simpósio Brasileiro de Sistemas Colaborativos*, pages 174–182, Porto Alegre, RS, Brasil. SBC.

Hemmat, A., Sharbaf, M., Kolahdouz-Rahimi, S., Lano, K., and Tehrani, S. Y. (2025). Research directions for using llm in software requirement engineering: a systematic review. *Frontiers in Computer Science*, Volume 7 - 2025:1519437.

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., and Wang, H. (2024). Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 33(8):1–79.

Huang, K., Wang, F., Huang, Y., and Arora, C. (2025). Prompt engineering for requirements engineering: A literature review and roadmap.

Kasauli, R., Knauss, E., Horkoff, J., Liebel, G., and de Oliveira Neto, F. G. (2021). Requirements engineering challenges and practices in large-scale agile system development. *Journal of Systems and Software*, 172:110851.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Mantau, M. and Benitti, F. (2023). The awareness assessment model: measuring the awareness and collaboration support over the participant's perspective. In *Anais do XVIII Simpósio Brasileiro de Sistemas Colaborativos*, pages 30–43, Porto Alegre, RS, Brasil. SBC.

Marques, A. B., Fiori, M. V., and Ferreira, B. (2023). Evolving the usarp method to support usability requirements elicitation and specification. In *Anais do Workshop em Engenharia de Requisitos - Proceedings of the 26th Workshop on Requirements Engineering (WER2023)*.

Quattrocchi, G., Pasquale, L., Spoletini, P., and Baresi, L. (2026). Can llms generate user stories and assess their quality? *IEEE Transactions on Software Engineering*, pages 1–17.

Silva, E. B., Andrade, M. E., Lima, A. M., and Marques, A. B. (2024). Unlock the power of the usarp method: The usarp tool contribution to usability requirements elicitation and specification. In *Anais do Workshop em Engenharia de Requisitos - Proceedings of the 27th Workshop on Requirements Engineering (WER2024)*.

Stefani, C. E. and Duduchi, M. (2023). Elementos de colaboração nos métodos ágeis de desenvolvimento de software. In *Anais do XVIII Simpósio Brasileiro de Sistemas Colaborativos*, pages 86–100, Porto Alegre, RS, Brasil. SBC.

Valadão, J., Andrade, J., and Cordeiro Neto, J. (2014). Abordagens sociotécnicas e os estudos em tecnologias sociais. *Pretexto*, 15:44–61.

Vogelsang, A. (2024). Prompting the future: Integrating generative llms and requirements engineering. In *REFSQ Workshops*.

Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2024). *Experimentation in Software Engineering*. Springer Berlin / Heidelberg, 2 edition.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). Prompting large language models for human-centered decision making. *ACM Transactions on Computer-Human Interaction*, 30(5):1–31.