

Design da Avaliação — SofIA (SBES 2026)

Documento consolidado — versão final (pós-coleta)

Artigo: *SofIA: Software Engineering Planning Tool with IA* - Trilha de Ferramentas, SBES 2026 (CBSOFT, São Paulo), Autores: Sidney Alex de Amorim Arruda e Vinicius Cardoso Garcia (CIn/UFPE)

1. Posicionamento Epistemológico e Escolha Metodológica

Tipo de contribuição e enquadramento

O SofIA é um artefato de software, uma ferramenta que resolve um problema prático (a ausência ou deficiência de artefatos de planejamento nas fases iniciais de projetos de software). Isso situa o trabalho no paradigma de **Design Science Research (DSR)**, conforme Hevner et al. (2004). No ciclo DSR, a construção do artefato já foi realizada (TCC, 2024); o que falta para completar o ciclo é a **avaliação do artefato** em contexto de uso. A Trilha de Ferramentas do SBES espera exatamente isso: demonstrar que o artefato existe, funciona e foi avaliado com algum rigor.

Tipo de estudo empírico

Considerando as restrições, prazo curto (execução até ~18 de maio), participantes voluntários recrutados por conveniência, uma única ferramenta sem baseline comparativo controlado, um experimento controlado está fora de alcance. Um estudo de caso no sentido de Yin (caso único em profundidade) seria viável mas renderia dados qualitativos difíceis de comunicar em 0,75–1 página. A opção mais adequada é um **survey baseado em tarefa** (o que Wohlin et al. classificam como *survey* com componente observacional): os participantes executam uma tarefa predefinida com a ferramenta e, em seguida, respondem a um questionário estruturado.

Dentro do ciclo DSR, a avaliação pode ser feita por diferentes métodos. Peffers et al. (2007), que operacionalizam o DSR em seis atividades, aceitam surveys e estudos de campo como formas legítimas de avaliação de artefato, especialmente quando a contribuição principal é a ferramenta em si, e não uma teoria. O survey baseado em tarefa permite coletar dados quantitativos sobre **usabilidade** (via SUS, Brooke, 1996) e **utilidade/intenção de uso** (via construtos do TAM, Davis, 1989), ambos amplamente aceitos na comunidade de ES e comparáveis com benchmarks publicados.

Estrutura metodológica

O framework GQM (Basili et al., 1994) é usado para derivar, de forma rastreável, os objetivos da avaliação → perguntas de pesquisa → métricas concretas.

Questão ética (CEP/CAAE)

Pesquisas que envolvem coleta de opinião de adultos voluntários sobre uma ferramenta, sem intervenção clínica ou manipulação de dados sensíveis, geralmente se enquadram na Resolução CNS 510/2016, Art. 1º, parágrafo único, inciso VII, que isenta de registro no CEP pesquisas de opinião pública com participantes não identificáveis. Como os dados serão anonimizados e não envolvem populações vulneráveis, a avaliação **dispensa submissão ao CEP**, mas inclui um Termo de Consentimento Livre e Esclarecido (TCLE) simplificado na abertura do formulário, informando objetivo, anonimização e voluntariedade.

Resumo da proposta metodológica

O estudo se posiciona como a **atividade de avaliação** dentro de um ciclo DSR. O método empírico é um **survey baseado em tarefa** (Wohlin et al., 2012), com instrumentos padronizados (SUS para usabilidade; construtos adaptados do TAM para utilidade percebida e intenção de uso). A estruturação dos objetivos segue o GQM. A execução é remota e assíncrona (participante baixa a ferramenta, executa a tarefa, responde o formulário online), com sessão síncrona opcional de 30 minutos via Google Meet para suporte à instalação/configuração.

2. Objetivos da Avaliação no Formato GQM

Cada objetivo segue o template canônico de Basili et al.: *Analisar [objeto] com o propósito de [intenção] com respeito a [foco de qualidade] do ponto de vista de [perspectiva] no contexto de [contexto]*.

G1 — Usabilidade

Analisar o SoflA **com o propósito de** avaliar **com respeito à** usabilidade percebida **do ponto de vista de** profissionais e estudantes avançados de engenharia de software **no contexto de** geração de artefatos de planejamento a partir de uma descrição textual de projeto.

Q1.1. Os usuários consideram a ferramenta fácil de aprender e operar?

- **M1.1.** Escore SUS (0–100), calculado conforme Brooke (1996). Benchmark de comparação: média global de 68 pontos (Sauro, 2011).

Q1.2. Quais aspectos da interface ou do fluxo de uso geram maior dificuldade?

- **M1.2.** Distribuição das respostas por item individual do SUS (itens com média $\leq 3,0$ na escala original de 5 pontos sinalizam pontos problemáticos).
- **M1.3.** Categorização temática das respostas à pergunta aberta sobre dificuldades encontradas (dado qualitativo complementar).

G2 — Utilidade Percebida dos Artefatos Gerados

Analisar os artefatos de planejamento produzidos pelo SoflA **com o propósito de** avaliar **com respeito à** utilidade percebida como ponto de partida para o desenvolvimento de software **do ponto de vista de** profissionais e estudantes avançados de engenharia de software **no contexto de** um cenário de projeto fictício executado durante a avaliação.

Q2.1. Os artefatos gerados são percebidos como úteis para iniciar o planejamento de um projeto real?

- **M2.1.** Média e distribuição das respostas ao bloco de itens adaptados do construto *Perceived Usefulness* (PU) do TAM (Davis, 1989), em escala Likert de 7 pontos.

Q2.2. Os artefatos gerados são percebidos como suficientemente completos e corretos para servir de base de trabalho?

- **M2.2.** Média e distribuição dos itens complementares desenvolvidos para esta avaliação sobre completude e correção dos artefatos, ancorados na mesma escala de 7 pontos do bloco TAM para manter consistência.

Q2.3. Quais tipos de artefato são percebidos como mais e menos úteis?

- **M2.3.** Ranking de utilidade percebida por tipo de artefato (prompt expandido, descrição/escopo do projeto, documento de requisitos, plano de projeto, documento de arquitetura, e mapas de tela com histórias de usuário e BDD), obtido por itens individuais de rating.

G3 — Intenção de Uso

Analisar o SoflA **com o propósito de** avaliar **com respeito à** intenção de adoção futura **do ponto de vista de** profissionais e estudantes avançados de engenharia de software **no contexto de** suas atividades profissionais ou acadêmicas.

Q3.1. Os participantes pretendem usar a ferramenta em projetos futuros?

- **M3.1.** Média e distribuição das respostas ao bloco de itens adaptados do construto *Behavioral Intention to Use* (BI) do TAM.

Q3.2. Que fatores influenciam positiva ou negativamente a intenção de uso?

- **M3.2.** Análise das respostas à pergunta aberta sobre fatores de adoção/rejeição (dado qualitativo).

Notas sobre decisões de design do GQM

- **Escala Likert de 7 pontos nos blocos TAM vs. 5 pontos no SUS.** O SUS usa obrigatoriamente 5 pontos, alterar invalida as normas de referência. O TAM original de Davis usa 7 pontos, e manter essa escala preserva a sensibilidade discriminativa do instrumento. A coexistência de duas escalas no mesmo formulário é prática comum em estudos de ES.
- **Perguntas abertas limitadas a duas.** Apenas duas questões abertas (M1.3 e M3.2) para não inflacionar o tempo de resposta e para que os dados qualitativos caibam no espaço do artigo como complemento.
- **Artefatos avaliados individualmente (M2.3).** Rating por tipo de artefato porque os dados do TCC sugerem que a percepção varia entre artefatos.

3. Design do Estudo

3.1 Tipo de estudo e justificativa

Survey baseado em tarefa, conforme posicionado na seção 1. Cada participante executa uma tarefa padronizada com o SofIA (gerar artefatos para um cenário de projeto predefinido), inspeciona os artefatos gerados e responde ao questionário. Não há grupo de controle, a avaliação mede percepção sobre o artefato (ferramenta), não compara tratamentos. Isso é consistente com a atividade de avaliação em DSR e com o escopo típico de artigos da Trilha de Ferramentas do SBES.

3.2 Perfil dos participantes e recrutamento

Crerérios de inclusão. Participantes devem satisfazer pelo menos uma das condições: (a) profissional atuante em desenvolvimento de software, engenharia de requisitos, arquitetura, gestão de projetos ou áreas correlatas, com no mínimo 1 ano de experiência; ou (b) estudante de graduação ou pós-graduação em Computação, Sistemas de Informação ou área afim, que já tenha cursado disciplina de Engenharia de Software ou equivalente.

Crerério de exclusão. Participantes que não completarem a tarefa (não gerarem ao menos um projeto completo no SofIA) terão suas respostas descartadas da análise.

Recrutamento. Amostragem por conveniência via rede de contatos acadêmica e profissional dos autores (colegas do CIn/UFPE, contatos do grupo de pesquisa do orientador, comunidades de desenvolvedores no LinkedIn e grupos de Telegram/Discord de ES). O convite será padronizado e conterá o TCLE

simplificado. O recrutamento por conveniência será declarado como limitação no artigo.

3.3 Tamanho amostral

O SUS possui benchmarks consolidados para amostras a partir de 12–15 participantes (Tullis & Stetson, 2004, demonstraram que com 12 respondentes os escores SUS se estabilizam com correlação $\geq 0,90$ em relação a amostras maiores). Para os construtos do TAM, amostras de 20–30 participantes permitem calcular médias e intervalos de confiança com razoabilidade em estudos exploratórios, embora não sustentem análise fatorial confirmatória, o que não é nosso objetivo.

Meta: 20 participantes válidos (respostas completas + tarefa executada). Para alcançar isso com a taxa de abandono típica de estudos remotos assíncronos (~25–30%), recrutaremos no mínimo 28–30 pessoas. O artigo reportará tanto o número de recrutados quanto o de respostas válidas.

Resultado efetivo: 13 participantes recrutados, todos completaram a tarefa e o questionário (n = 13 válidos, taxa de conclusão de 100%). A amostra ficou abaixo da meta de 20, mas acima do limiar de estabilização do SUS (12–15). Essa limitação é declarada na seção de ameaças à validade.

3.4 Procedimento

O participante recebe, por e-mail ou mensagem, um pacote contendo:

(a) Termo de Consentimento Livre e Esclarecido (TCLE), embutido como primeira seção do formulário online. O participante só avança se concordar.

(b) Instruções de instalação e configuração, documento curto (1–2 páginas) com passo a passo para instalar o SoflA Desktop e configurar a conexão com o modelo de linguagem. O caminho padrão recomendado é o **Granite 4.0 H Tiny** (arquitetura GraniteHybrid, 7B parâmetros, quantização GGUF Q4_K_M, aproximadamente 4,23 GB, licença Apache 2.0) via **LM Studio**. O LM Studio oferece uma interface gráfica para download e execução de modelos locais, dispensando linha de comando — o que reduz a fricção para participantes menos familiarizados com ferramentas de terminal. O requisito mínimo de hardware sobe em relação a modelos menores: recomendamos 16 GB de RAM e GPU com ao menos 6 GB de VRAM para execução fluida, ou 8 GB de RAM com execução em CPU (mais lenta, mas funcional). Participantes que preferirem usar API comercial (OpenAI, Anthropic, Gemini, OpenRouter) ou outro modelo local podem fazê-lo, o questionário registra a escolha. A padronização pelo Granite 4.0 H Tiny visa maximizar a homogeneidade dos artefatos avaliados e, consequentemente, a comparabilidade das respostas do bloco G2.

Resultado efetivo: todos os 13 participantes utilizaram o Granite 4.0 H Tiny via LM Studio conforme recomendado, homogeneidade total no modelo utilizado.

(c) Cenário da tarefa: descrição de um projeto fictício padronizado que o participante usará como entrada no SoflA. Todos usam o mesmo cenário para permitir comparabilidade dos artefatos gerados.

(d) Roteiro da tarefa: sequência de passos: instalar o SoflA e o LM Studio, baixar o modelo Granite 4.0 H Tiny, iniciar o servidor local, configurar o modelo no SoflA, criar o projeto com o cenário fornecido, aguardar a geração completa dos seis artefatos (prompt expandido, descrição/escopo, requisitos, plano de projeto, arquitetura, mapas de tela com histórias de usuário e BDD), inspecionar cada artefato dedicando ao menos 2–3 minutos por tipo, registrar o tempo total.

(e) Sessão síncrona opcional: participantes que encontrarem dificuldades na instalação ou configuração podem agendar uma sessão de até 30 minutos via Google Meet com um dos autores. A sessão se limita a suporte técnico (instalação, configuração do LM Studio, conexão com o SoflA) e não interfere na execução da tarefa nem no preenchimento do questionário.

(f) Questionário online (Google Forms ou equivalente): respondido após a tarefa. Estrutura: dados demográficos → SUS → bloco TAM-PU → itens complementares de completude/correção → rating por artefato → bloco TAM-BI → perguntas abertas.

Tempo estimado total: 45–70 minutos (15–25 min instalação/configuração com LM Studio + download do modelo de ~4,23 GB + 15–20 min geração e inspeção dos artefatos + 10–15 min questionário). O download do modelo é o fator mais variável: em conexões de 10 Mbps leva aproximadamente 6 minutos; em conexões mais lentas pode ultrapassar 15 minutos. O roteiro orientará o participante a iniciar o download antes de ler o restante das instruções para paralelizar o tempo de espera.

Resultado efetivo: o tempo total ficou acima do estimado: 9 participantes reportaram 60–90 minutos, e 4 ultrapassaram 90 minutos. Nenhum participante completou em menos de 60 minutos.

3.5 Controle de variáveis de confusão

Dado que não é um experimento controlado, não "controlamos" variáveis no sentido estrito, mas adotamos estratégias para mitigar as principais ameaças:

Variabilidade do LLM utilizado. Participantes diferentes podem usar modelos diferentes, o que afeta a qualidade dos artefatos e a percepção de utilidade. *Mitigação:* a recomendação padronizada do Granite 4.0 H Tiny via LM Studio deve

concentrar a maioria dos participantes em um único modelo. Diferentemente de modelos menores (~700 MB), o Granite 4.0 H Tiny exige hardware mais capaz (recomendados 16 GB de RAM e GPU com 6 GB de VRAM), o que pode excluir participantes com máquinas mais modestas, nesses casos, a alternativa é execução em CPU (mais lenta) ou uso de API comercial com chave própria. O questionário (P2.1) registra qual modelo/provedor foi usado. Na análise, se houver participantes suficientes usando modelos alternativos, compararemos os escores de G2 entre subgrupos; caso contrário, reportaremos a proporção que seguiu o caminho padrão e declararemos a limitação. A barreira de hardware mais elevada será declarada como possível fonte de viés de seleção (participantes com máquinas potentes podem ter perfil profissional distinto).

Resultado efetivo: todos os 13 participantes usaram o modelo recomendado. A variável de modelo LLM foi completamente controlada nesta amostra.

Variabilidade de experiência dos participantes. Participantes mais experientes podem avaliar os artefatos com mais rigor. *Mitigação:* coletamos anos de experiência e papel profissional no bloco demográfico (P1.2, P1.3), permitindo análise estratificada ou, no mínimo, descrição da composição da amostra.

Cenário padronizado vs. projeto real. Usar um cenário fictício garante comparabilidade, mas reduz o realismo, o participante não está resolvendo um problema próprio. *Mitigação:* declaramos essa trade-off como limitação.

Efeito de novidade. O participante está usando a ferramenta pela primeira vez, o que pode inflacionar positivamente (entusiasmo) ou negativamente (frustração com configuração) as respostas. *Mitigação:* o SUS já é desenhado para capturar impressões de primeiro uso; declaramos a limitação de que os resultados refletem uma sessão única.

Viés de seleção. Amostragem por conveniência tende a recrutar participantes com interesse prévio em IA ou em ferramentas de planejamento, o que pode enviesar positivamente os resultados. *Mitigação:* coletamos no bloco demográfico a familiaridade prévia com ferramentas de IA para ES (P1.5) e declaramos o viés como limitação.

3.6 Variáveis coletadas — resumo

Variável	Tipo	Fonte
Experiência profissional (anos, papel)	Demográfica	Questionário (P1.2, P1.3)
Familiaridade com ferramentas de IA para ES	Demográfica	Questionário (P1.5)
Hábito de documentação de requisitos	Demográfica	Questionário (P1.4)

Variável	Tipo	Fonte
Modelo/provedor LLM utilizado	Contextual	Questionário (P2.1)
Tempo total da tarefa	Contextual	Questionário (P2.2)
Uso da sessão síncrona de suporte	Contextual	Questionário (P2.3)
Escore SUS	Dependente (G1)	Questionário, 10 itens SUS
Utilidade percebida (PU)	Dependente (G2)	Questionário, itens TAM-PU
Completeness, consistência, correção e aplicabilidade	Dependente (G2)	Questionário, itens complementares (CC1-CC4)
Utilidade por tipo de artefato	Dependente (G2)	Questionário, rating individual (AT1-AT6)
Intenção de uso (BI)	Dependente (G3)	Questionário, itens TAM-BI
Dificuldades encontradas	Qualitativa	Pergunta aberta (PA1)
Fatores de adoção/rejeição	Qualitativa	Pergunta aberta (PA2)

4. Instrumento de Coleta — Questionário Completo

O questionário está organizado em sete seções. As instruções entre colchetes são notas internas, não aparecem para o participante.

SEÇÃO 0 — Termo de Consentimento Livre e Esclarecido

[Primeira tela do formulário. O participante só avança se marcar "Concordo".]

Texto do TCLE:

Você está sendo convidado(a) a participar de uma avaliação da ferramenta SofIA (Software Engineering Planning Tool with IA), conduzida por pesquisadores do Centro de Informática da Universidade Federal de Pernambuco (CIn/UFPE). A participação consiste em usar a ferramenta para gerar artefatos de planejamento de software a partir de um cenário fornecido e, em seguida, responder a este questionário. O tempo estimado é de 45 a 70 minutos.

Sua participação é voluntária. Você pode desistir a qualquer momento sem necessidade de justificativa. As respostas são anônimas, nenhum dado que permita sua identificação pessoal será coletado ou publicado. Os resultados serão utilizados exclusivamente para fins de pesquisa acadêmica.

Em caso de dúvidas, entre em contato com os pesquisadores:
saaa@cin.ufpe.br / vcg@cin.ufpe.br.

☐ Li e concordo em participar desta avaliação.

SEÇÃO 1 — Perfil do Participante

[Objetivo: caracterizar a amostra e permitir análise estratificada. Todas as perguntas são obrigatórias.]

P1.1. Qual é sua formação acadêmica atual ou mais recente?

- Graduação em andamento
- Graduação concluída
- Pós-graduação (especialização/MBA)
- Mestrado (em andamento ou concluído)
- Doutorado (em andamento ou concluído)

P1.2. Qual é o seu papel profissional principal atualmente?

- Desenvolvedor(a) de software
- Engenheiro(a) / Analista de requisitos
- Arquiteto(a) de software
- Tech Lead / Líder técnico
- Gerente de projetos / Scrum Master / PO
- Estudante (sem atuação profissional na área)
- Outro: _____

P1.3. Há quanto tempo você atua profissionalmente na área de desenvolvimento de software?

- Ainda não atuo profissionalmente
- Menos de 1 ano
- 1 a 3 anos
- 4 a 7 anos
- 8 a 15 anos
- Mais de 15 anos

P1.4. Com que frequência você elabora ou utiliza artefatos de planejamento de software (requisitos, arquitetura, plano de projeto, histórias de usuário) no seu trabalho ou estudos?

- Nunca
- Raramente (em poucos projetos)
- Ocasionalmente (em alguns projetos)

- Frequentemente (na maioria dos projetos)
- Sempre (em todos os projetos)

P1.5. Antes desta avaliação, você já havia utilizado ferramentas que empregam IA/LLMs especificamente para gerar artefatos de planejamento de software (excluindo uso genérico de ChatGPT ou similares para tirar dúvidas)?

- Nunca utilizei
- Já experimentei, mas não uso regularmente
- Utilizo regularmente

SEÇÃO 2 — Contexto da Tarefa Realizada

[Objetivo: registrar variáveis contextuais que podem influenciar as respostas.]

P2.1. Qual modelo de linguagem (LLM) você utilizou durante a avaliação?

- Granite 4.0 H Tiny via LM Studio (modelo recomendado)
- Outro modelo via LM Studio: _____
- Outro modelo via Ollama: _____
- API da OpenAI (modelo: _____)
- API da Anthropic (modelo: _____)
- API do Google Gemini (modelo: _____)
- OpenRouter (modelo: _____)

P2.2. Aproximadamente, quanto tempo levou desde o início da instalação até a conclusão da inspeção de todos os artefatos gerados?

- Menos de 30 minutos
- 30 a 45 minutos
- 45 a 60 minutos
- 60 a 90 minutos
- Mais de 90 minutos

P2.3. Você precisou utilizar a sessão de suporte via Google Meet?

- Não
- Sim

SEÇÃO 3 — Usabilidade (SUS — System Usability Scale)

[Instrumento padronizado de Brooke (1996). 10 itens, escala de 5 pontos. Os itens são traduzidos para português seguindo a tradução validada de Tenório et al. (2011). A ordem e a alternância positivo/negativo são obrigatórias e não devem ser alteradas.]

Instrução ao participante: Para cada afirmação abaixo, indique o grau em que você concorda ou discorda, considerando sua experiência ao usar o SofIA. (1 = Discordo totalmente, 5 = Concordo totalmente)

#	Item	1	2	3	4	5
S1	Eu gostaria de usar o SofIA com frequência.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S2	Achei o SofIA desnecessariamente complexo.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S3	Achei o SofIA fácil de usar.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S4	Acho que precisaria de apoio técnico para conseguir usar o SofIA.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S5	Achei que as diversas funções do SofIA estavam bem integradas.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S6	Achei que havia muita inconsistência no SofIA.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S7	Imagino que a maioria das pessoas aprenderia a usar o SofIA rapidamente.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S8	Achei o SofIA muito complicado de usar.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S9	Me senti confiante ao usar o SofIA.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
S10	Precisei aprender muitas coisas antes de conseguir usar o SofIA.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Cálculo do escore: para itens ímpares (positivos), subtrair 1 da resposta; para itens pares (negativos), subtrair a resposta de 5. Somar tudo e multiplicar por 2,5. Resultado: 0–100.]

SEÇÃO 4 — Utilidade Percebida (Construtos TAM — Perceived Usefulness)

[Itens adaptados de Davis (1989). Escala Likert de 7 pontos.]

Instrução ao participante: Considerando os artefatos que o SoflA gerou a partir do cenário fornecido, indique o grau em que você concorda com cada afirmação. (1 = Discordo totalmente, 7 = Concordo totalmente)

PU1. Usar o SoflA me permitiria gerar artefatos de planejamento de software mais rapidamente do que fazê-los manualmente.

PU2. Os artefatos gerados pelo SoflA seriam úteis como ponto de partida para o planejamento de um projeto real.

PU3. Usar o SoflA melhoraria minha produtividade na fase de planejamento de software.

PU4. Usar o SoflA facilitaria a tarefa de elaborar a documentação inicial de um projeto de software.

[4 itens — reduzido do original de 6 itens de Davis para manter o questionário conciso. Os itens selecionados cobrem velocidade, utilidade, produtividade e facilidade, que são as facetas centrais do construto PU.]

[4 itens opcionais em formato texto aberto para a pessoa poder, caso se sinta confortável, expressar sua opinião sobre cada afirmação. Estas perguntas vêm logo após a pergunta em escala.]

PU1.1. Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre usar o SoflA te permitiria gerar artefatos de planejamento de software mais rapidamente do que fazê-los manualmente?

PU2.1. Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se os artefatos gerados pelo SoflA seriam úteis como ponto de partida para o planejamento de um projeto real?

PU3.1. Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se usar o SoflA melhoraria sua produtividade na fase de planejamento de software?

PU4.1. *Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se usar o SoflA facilitaria a tarefa de elaborar a documentação inicial de um projeto de software?*

SEÇÃO 5 — Completude e Correção dos Artefatos + Rating por Tipo

[Itens complementares desenvolvidos para esta avaliação. Mesma escala de 7 pontos para consistência.]

Instrução ao participante: Ainda sobre os artefatos gerados pelo SoflA, avalie os seguintes aspectos. (1 = Discordo totalmente, 7 = Concordo totalmente)

CC1. Os artefatos gerados cobrem os principais aspectos que eu esperaria encontrar em um planejamento inicial de software (completude).

CC2. As informações contidas nos artefatos são coerentes entre si — não há contradições evidentes entre os diferentes documentos gerados (consistência).

CC3. Os artefatos gerados contêm informações tecnicamente corretas e razoáveis para o cenário descrito (correção).

CC4. Os artefatos gerados precisariam de pouca edição para serem usados como documentação real de um projeto (aplicabilidade prática).

[4 itens opcionais em formato texto aberto para a pessoa poder, caso se sinta confortável, expressar sua opinião sobre cada afirmação. Estas perguntas vêm logo após a pergunta em escala.]

CC1.1. *Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se os artefatos gerados cobrem os principais aspectos que você esperaria encontrar em um planejamento inicial de software (completude)?*

CC2.1. *Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se as informações contidas nos artefatos são coerentes entre si, não há contradições evidentes entre os diferentes documentos gerados (consistência)?*

CC3.1. *Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se os artefatos gerados contêm informações tecnicamente corretas e razoáveis para o cenário descrito (correção)?*

CC4.1. *Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se os artefatos gerados precisariam de pouca edição para serem usados como documentação real de um projeto (aplicabilidade prática)?*

Instrução ao participante: Agora, avalie individualmente a utilidade de cada tipo de artefato gerado pelo SofIA. (1 = Nada útil, 7 = Extremamente útil)

Artefato	1	2	3	4	5	6	7
AT1. Contexto aprimorado (descrição o enriquecida do projeto)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AT2. Descrição / Escopo do projeto	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AT3. Documento de requisitos (funcionais e não funcionais)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AT4. Plano de projeto	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AT5. Documento de arquitetura	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AT6. Mapas de tela com histórias de usuário e BDD	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

SEÇÃO 6 — Intenção de Uso (Construto TAM — Behavioral Intention)

[Itens adaptados de Venkatesh et al. (2003) / Davis (1989). Escala de 7 pontos.]

Instrução ao participante: Considerando sua experiência com o SoflA, indique o grau em que concorda com as afirmações a seguir. (1 = Discordo totalmente, 7 = Concordo totalmente)

BI1. Eu pretendo usar o SoflA em projetos futuros (acadêmicos ou profissionais).

BI2. Eu recomendaria o SoflA para colegas que trabalham com desenvolvimento de software.

BI3. Se disponível, eu incorporaria o SoflA ao meu fluxo de trabalho na fase de planejamento.

[3 itens opcionais em formato texto aberto para a pessoa poder, caso se sinta confortável, expressar sua opinião sobre cada afirmação. Estas perguntas vêm logo após a pergunta em escala.]

BI1.1. Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se você pretende usar o SoflA em projetos futuros (acadêmicos ou profissionais)?

BI2.1. Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se você recomendaria o SoflA para colegas que trabalham com desenvolvimento de software?

BI3.1. Caso sinta-se à vontade, qual a justificativa para a resposta anterior sobre se disponível, você incorporaria o SoflA ao seu fluxo de trabalho na fase de planejamento?

SEÇÃO 7 — Perguntas Abertas

PA1. Descreva brevemente as principais dificuldades que você encontrou ao usar o SoflA (instalação, configuração, uso da interface, compreensão dos artefatos ou qualquer outro aspecto). Se não encontrou dificuldades, escreva "Nenhuma".

PA2. Que fatores mais influenciariam sua decisão de adotar ou não o SoflA no seu dia a dia? (Pode mencionar aspectos positivos que incentivam a adoção e aspectos negativos que desincentivariam.)

Resumo do instrumento

Seção	Itens	Tempo estimado
0 — TCLE	1	< 1 min
1 — Perfil	5	2 min

Seção	Itens	Tempo estimado
2 — Contexto	3	1 min
3 — SUS	10	3 min
4 — TAM-PU	4	2 min
5 — Completude/Correção + Rating	4 + 6	3 min
6 — TAM-BI	3	1 min
7 — Abertas	2	3–5 min
Total	38 itens	~15 min

Notas sobre decisões do instrumento

- Os itens PU foram reduzidos de 6 para 4 e os itens BI de 4 para 3 em relação aos originais de Davis. A redução é intencional para manter o questionário abaixo de 15 minutos. Os itens removidos tinham alta correlação com os mantidos nos estudos de validação originais.
- A tradução do SUS segue a versão brasileira validada por Tenório et al. (2011).
- Os itens CC1-CC4 serão declarados no artigo como "itens complementares desenvolvidos para esta avaliação". Não possuem validação prévia, isso será registrado como limitação.

5. Protocolo da Tarefa

5.1 Cenário fictício padronizado

O cenário atende a quatro critérios: (i) compreensível por qualquer participante do perfil definido; (ii) complexidade suficiente para artefatos com substância avaliável; (iii) não tão complexo a ponto de dificultar a avaliação por participantes menos experientes; (iv) domínio familiar o bastante para julgar se os artefatos fazem sentido.

Cenário utilizado:

Nome do projeto: TaskFlow

Descrição para inserir no SofIA:

"Preciso de uma aplicação web para gerenciamento de projetos voltada a equipes pequenas de desenvolvimento de software (3 a 10 pessoas). A aplicação deve permitir criar projetos, organizar tarefas em quadros Kanban com colunas personalizáveis, atribuir responsáveis e prazos a cada tarefa, e acompanhar o progresso por meio de um dashboard com métricas básicas (tarefas concluídas por sprint, tarefas atrasadas, distribuição de carga por membro). Deve haver um sistema de notificações por e-mail quando uma tarefa for atribuída ou estiver próxima do prazo. O acesso será controlado por autenticação com e-mail e senha, com dois perfis: administrador do projeto e membro da equipe."

Justificativa. Gerenciamento de projetos com Kanban é um domínio que praticamente todo profissional ou estudante avançado de ES conhece por experiência direta. A descrição tem ~100 palavras, com riqueza suficiente (perfis de usuário, funcionalidade transversal de notificações, métricas esperadas) para gerar artefatos com profundidade, enquanto deixa lacunas naturais (tecnologias, integrações externas, requisitos não funcionais de performance) que a ferramenta deverá preencher.

5.2 Roteiro do participante

AVALIAÇÃO DO SofIA — Roteiro da Tarefa

Tempo estimado: 45–70 minutos no total.

Leia este roteiro inteiro antes de começar.

ETAPA 1 — Instalação (15–25 min)

1. Baixe e instale o SofIA Desktop a partir do link fornecido. Versões disponíveis para Windows, macOS e Linux.
2. Baixe e instale o **LM Studio** a partir de <https://lmstudio.ai>. Siga as instruções padrão para seu sistema operacional. O LM Studio possui interface gráfica, não é necessário usar terminal.
3. Abra o LM Studio. Na barra de busca, pesquise por **granite-4.0-h-tiny**. Selecione a versão **GGUF Q4_K_M** (~4,23 GB) e clique em **Download**. Aguarde a conclusão do download. **Dica:** enquanto o download ocorre, aproveite para ler o restante deste roteiro.
4. Após o download, vá à aba **Local Server** (ícone de servidor na barra lateral do LM Studio). Selecione o modelo **granite-4.0-h-tiny** e clique em **Start**

Server. Aguarde até que o status indique que o servidor está ativo (porta padrão: 1234).

5. Abra o SoflA Desktop. Na tela de configuração, selecione **LM Studio** como provedor. O endereço padrão (<http://localhost:1234>) já deve estar preenchido. Salve a configuração.

Se encontrar qualquer dificuldade nesta etapa, agende uma sessão de suporte de até 30 minutos via Google Meet pelo link fornecido.

ETAPA 2 — Criação do projeto e geração dos artefatos (15–20 min)

6. Na tela inicial do SoflA, clique em **Criar Projeto**.
7. No campo **Título**, insira: **TaskFlow**
8. No campo **Informações**, copie e cole exatamente o texto do cenário fornecido.
9. Clique em **Criar Projeto** e aguarde a geração de todos os artefatos. O SoflA exibirá o progresso na tela de status. Com o modelo Granite 4.0 H Tiny, o processo completo costuma levar entre 5 e 20 minutos dependendo do hardware (GPU dedicada vs. CPU). **Não feche a aplicação nem o LM Studio durante a geração.**

ETAPA 3 — Inspeção dos artefatos (10–15 min)

10. Após a conclusão da geração, navegue por **cada um** dos seguintes artefatos e leia o conteúdo com atenção:
 - Contexto aprimorado (descrição enriquecida do projeto)
 - Descrição / Escopo do projeto
 - Documento de requisitos (funcionais e não funcionais)
 - Plano de projeto
 - Documento de arquitetura
 - Mapas de tela com histórias de usuário e BDD

Dedique ao menos **2 a 3 minutos por artefato**. Ao ler, considere mentalmente: este artefato seria útil como ponto de partida se eu fosse de fato desenvolver o TaskFlow? O que está bom? O que está faltando ou parece incorreto?

11. Anote o horário de término da inspeção. Você usará essa informação no questionário.

ETAPA 4 — Questionário (10–15 min)

12. Acesse o questionário pelo link fornecido.
13. Responda todas as seções com base na experiência que acabou de ter. Não há respostas certas ou erradas, queremos sua percepção honesta.
14. Ao finalizar, clique em **Enviar**. Pronto, agradecemos muito pela participação!

5.3 Notas operacionais

Teste piloto. Antes de abrir o recrutamento, foram executados testes piloto com o roteiro completo em pelo menos duas máquinas com perfis diferentes (uma com GPU dedicada ≥ 6 GB VRAM, outra com CPU apenas) e cronometrado cada etapa.

Janela de contexto. A janela de contexto do Granite 4.0 H Tiny foi validada no piloto para confirmar que comporta a geração sequencial dos seis artefatos sem truncamento.

Versionamento do cenário. O texto do cenário não pode ser alterado após o início do recrutamento.

Registro de incidentes. Se participantes reportarem falhas na geração (artefato incompleto, erro da ferramenta), registrar o incidente e decidir caso a caso se a resposta é válida.

6. Plano de Análise

6.1 Preparação dos dados

Após o encerramento da coleta, os dados brutos do formulário serão exportados em CSV. A limpeza incluirá: remoção de respostas incompletas ou inválidas, verificação de respostas padrão suspeitas (todas as alternativas idênticas em uma seção inteira), e codificação das variáveis categóricas.

6.2 Análise de G1 — Usabilidade (SUS)

Cálculo do escore SUS. Para cada respondente, aplicar a fórmula padrão de Brooke (1996): itens ímpares (positivos) contribuem com [resposta - 1], itens pares (negativos) contribuem com [5 - resposta]; soma dos 10 valores multiplicada por 2,5, gerando escore de 0 a 100.

Estatísticas descritivas. Média, mediana, desvio padrão e intervalo de confiança de 95% do escore SUS. Classificação na escala adjetiva de Bangor et al. (2009): < 51

= ruim, 51–68 = ok, 68–80,3 = bom, > 80,3 = excelente. Comparação com o benchmark global de 68 pontos (Sauro, 2011).

Análise por item. Média de cada item individual (na escala original de 1–5). Itens com média $\leq 2,5$ (para positivos) ou $\geq 3,5$ (para negativos) serão sinalizados como pontos de atenção na discussão.

Análise exploratória de subgrupos. Teste de Mann-Whitney para comparar escores SUS entre participantes com mais vs. menos experiência profissional (ponto de corte: 4 anos). Se a amostra permitir ($n \geq 10$ por subgrupo), reportar; caso contrário, apenas descrever tendências sem teste inferencial.

6.3 Análise de G2 — Utilidade Percebida

Escore agregado PU. Média dos 4 itens PU1–PU4 por respondente. Reportar média geral, mediana, desvio padrão e IC 95%. Verificar consistência interna via alpha de Cronbach — espera-se $\alpha \geq 0,70$.

Itens complementares CC1–CC4. Reportar média e distribuição de cada item individualmente. Como são itens autorais sem validação prévia, não calcular escore agregado.

Rating por artefato (AT1–AT6). Reportar média e desvio padrão por tipo de artefato. Ordenar do mais ao menos bem avaliado. Teste de Friedman (alternativa não paramétrica à ANOVA de medidas repetidas). Se significativo ($p < 0,05$), testes post-hoc de Nemenyi.

Análise exploratória por modelo LLM. Se houver subgrupo suficiente ($n \geq 5$) de participantes que não usaram o Granite 4.0 H Tiny, comparar escores PU e ratings AT1–AT6 via Mann-Whitney.

6.4 Análise de G3 — Intenção de Uso

Escore agregado BI. Média dos 3 itens BI1–BI3 por respondente. Reportar média geral, mediana, desvio padrão e IC 95%. Verificar alpha de Cronbach.

Correlação PU × BI. Correlação de Spearman entre escore PU médio e escore BI médio por participante.

Correlação SUS × BI. Idem, para verificar se a usabilidade percebida também se associa à intenção de uso.

6.5 Análise qualitativa (PA1 e PA2)

Método. Análise temática simplificada (Braun & Clarke, 2006) em duas rodadas: (1) codificação aberta por um dos autores; (2) agrupamento dos códigos em temas. Temas reportados com frequência de menções e exemplos anonimizados.

Integração com dados quantitativos. Os temas de PA1 serão cruzados com os itens problemáticos do SUS. Os temas de PA2 serão discutidos à luz dos escores PU e BI.

6.6 Visualizações planejadas para o artigo

- **Tabela 1 — Perfil da amostra.** Distribuição compacta: uma linha por categoria, colunas n e %.
- **Figura 1 — Escore SUS.** Boxplot com linha de referência em 68 pontos. Alternativa: apenas texto.
- **Figura 2 — Rating por artefato (AT1-AT6).** Barras horizontais ordenadas por média, com barras de erro IC 95%.
- **No texto:** médias PU, CC1-CC4, BI com IC 95%; correlações com ρ e p-valor; temas qualitativos em prosa.

6.7 Ferramentas de análise

Todas as análises em **Python** (pandas, scipy.stats, pingouin, matplotlib/seaborn) ou **R** (psych, ggplot2).

6.8 Resumo das análises por objetivo

Objetivo	Métrica principal	Teste / Procedimento	Visualização
G1 - Usabilidade	Escore SUS	Descritivas + Bangor + Mann-Whitney	Boxplot ou texto
G2 - Utilidade (PU)	Média PU1-PU4	Descritivas + Cronbach α	Texto
G2 - Completude / Correção	CC1-CC4 individuais	Descritivas por item	Texto
G2 - Por artefato	AT1-AT6	Friedman + Nemenyi	Barras horizontais
G2 - Efeito do modelo	PU e AT por subgrupo	Mann-Whitney (se n suficiente)	Texto
G3 - Intenção de uso	Média BI1-BI3	Descritivas + Cronbach α	Texto
G3 × G2	PU × BI	Spearman ρ	Texto
G1 × G3	SUS × BI	Spearman ρ	Texto

Objetivo	Métrica principal	Teste / Procedimento	Visualização
Qualitativo	Temas PA1, PA2	Análise temática simplificada	Prosa

7. Resultados da Avaliação

7.1 Perfil da amostra (n = 13)

Dos 13 participantes recrutados, todos completaram a tarefa e o questionário (taxa de conclusão de 100%). A totalidade utilizou o modelo Granite 4.0 H Tiny via LM Studio conforme recomendado. Nove participantes reportaram tempo total entre 60 e 90 minutos; os quatro restantes ultrapassaram 90 minutos.

Característica	n	%
Formação acadêmica		
Graduação em andamento	2	15,4
Graduação concluída	6	46,2
Pós-graduação (espec./MBA)	4	30,8
Mestrado	1	7,7
Papel profissional		
Desenvolvedor(a)	5	38,5
Analista de requisitos	2	15,4
Tech Lead	2	15,4
PO / Scrum Master	2	15,4
Arquiteto(a) de software	1	7,7
Estudante	1	7,7
Experiência profissional		
Nenhuma	1	7,7
Menos de 1 ano	1	7,7

Característica	n	%
1 a 3 anos	2	15,4
4 a 7 anos	5	38,5
8 a 15 anos	3	23,1
Mais de 15 anos	1	7,7
Familiaridade com IA para ES		
Nunca utilizou	4	30,8
Já experimentou	5	38,5
Utiliza regularmente	4	30,8

A amostra é predominantemente composta por profissionais atuantes: 69,2% possuem quatro ou mais anos de experiência, e os papéis representados cobrem desde desenvolvimento até gestão de projetos. A distribuição de familiaridade com IA para ES é equilibrada, aproximadamente um terço por faixa.

7.2 Usabilidade (G1)

Métrica	Valor
Escore SUS médio	68,65
Desvio padrão	6,74
IC 95%	[64,58; 72,73]
Classificação Bangor	"Bom"
Relação ao benchmark (68)	Ligeiramente acima

Itens de atenção (itens negativos, médias altas indicam concordância com afirmação desfavorável):

- S4 ("Acho que precisaria de apoio técnico para conseguir usar o SoflA"): M = 2,54
- S6 ("Achei que havia muita inconsistência no SoflA"): M = 2,92

7.3 Utilidade percebida (G2)

Bloco TAM-PU (escala de 7 pontos):

Item	Descrição	Média
PU1	Ganho de velocidade	6,08
PU2	Ponto de partida real	5,08
PU3	Ganho de produtividade	5,31
PU4	Facilidade na documentação	5,92
Agregado		5,60 (DP = 0,52; IC 95%: [5,28; 5,91]; $\alpha = 0,82$)

Itens complementares (escala de 7 pontos):

Item	Descrição	Média
CC1	Compleitude	5,15
CC2	Consistência entre artefatos	4,15
CC3	Correção técnica	4,15
CC4	Aplicabilidade prática (pouca edição)	3,00

Rating por artefato (n = 11; escala de 7 pontos):

Artefato	Média	Posição
AT3 — Documento de requisitos	5,00	1º (empatado)
AT6 — Mapas de tela / histórias / BDD	5,00	1º (empatado)
AT2 — Descrição / Escopo	4,36	3º (empatado)
AT1 — Prompt expandido	4,36	3º (empatado)
AT4 — Plano de projeto	4,09	5º
AT5 — Documento de arquitetura	3,45	6º

Teste de Friedman: $\chi^2 = 20,61$, $p = 0,0010$ (significativo). **Post-hoc de Nemenyi:** AT3 vs AT5: $p = 0,0045$; AT6 vs AT5: $p = 0,0045$. Demais pares não significativos.

7.4 Intenção de uso (G3)

Item	Descrição	Média
BI1	Intenção de uso futuro	5,15
BI2	Recomendação a colegas	4,62
BI3	Incorporação ao fluxo de trabalho	5,15
Agregado		4,97 (DP = 0,71; IC 95%: [4,54; 5,41]; $\alpha = 0,89$)

Correlações:

Par	Spearman ρ	p-valor	Significância
PU \times BI	0,817	0,0007	Significativo
SUS \times BI	-0,269	0,374	Não significativo

A utilidade percebida é forte preditora da intenção de adoção (conforme previsto pelo TAM). A usabilidade da interface não se associou significativamente à intenção de uso nesta amostra, a percepção sobre a qualidade dos artefatos pesa mais na decisão de adoção do que a facilidade de uso da ferramenta.

7.5 Dados qualitativos

PA1 — Dificuldades (13 respondentes, 5 temas)

T1. Inflação de escopo não controlada (12/13). Funcionalidades não solicitadas introduzidas no contexto aprimorado propagam-se para todos os artefatos subsequentes. Participantes descreveram o fenômeno como "scope creep silencioso" e "contaminação". É a dificuldade mais frequente e a que mais afeta a percepção de aplicabilidade prática.

T2. Inconsistências internas entre artefatos (10/13). ORMs conflitantes, contagens divergentes de épicos, renomeação do projeto entre documentos e contradições entre escopo e arquitetura. Reflete a ausência de validação cruzada entre documentos na cadeia de geração.

Cruzamento com dados quantitativos: T1 e T2 corroboram as avaliações intermediárias de consistência (CC2 = 4,15) e, sobretudo, a baixa aplicabilidade prática (CC4 = 3,00). A propagação de escopo inflado e as inconsistências tornam os artefatos menos aproveitáveis sem edição significativa.

T3. Arquitetura superdimensionada (11/13). Proposição de microserviços + Kubernetes para um sistema destinado a equipes de 3 a 10 pessoas. Viés para padrões *enterprise* independente do porte descrito. Ausência de *Architecture Decision Records* (ADRs) e notação formal.

Cruzamento com dados quantitativos: T3 é consistente com a posição mais baixa do documento de arquitetura no ranking de artefatos (AT5 = 3,45) e com a diferença significativa em relação aos artefatos melhor avaliados ($p = 0,0045$).

T4. Números fabricados sem rastreabilidade (8/13). Orçamento, MTBF, latências e datas apresentados como estimativas sem derivação identificável. Os participantes não conseguem distinguir valores calculados de valores inventados.

T5. Problemas de formatação (5/13). Falhas de *encoding* em português, BDD em JSON bruto e HTML misturado com Markdown.

PA2 — Fatores de adoção (13 respondentes, 5 temas)

TA. Velocidade e cobertura como incentivos (13/13). Unanimidade. A geração de seis artefatos a partir de um parágrafo em poucos minutos é o diferencial principal reconhecido por todos os participantes.

TB. Operação local/stand-alone como fator decisivo (7/13). Diferencial frente a ChatGPT/Claude para projetos sensíveis em contextos corporativos com NDA ou restrições de privacidade.

TC. Integração com Jira/Confluence como barreira (6/13). Concentrado em POs, Scrum Masters e Tech Leads. O modelo *stand-alone* é percebido como incompatível com fluxos de trabalho centrados no ecossistema Atlassian.

TD. Demanda por controle de escopo e calibração de complexidade (11/13). Participantes solicitaram: parâmetro de agressividade na expansão do contexto, campo para indicar porte do projeto, distinção visual entre o prompt original e as inferências adicionadas pela ferramenta.

TE. BDD e épicos como artefatos mais aplicáveis (8/13). Output mais direto para importação em ferramentas de gestão com edição mínima. Consistente com a nota alta de AT6 ($M = 5,00$) na avaliação quantitativa.