Exploring LLMs on Supporting the Elicitation of Knowledge for NFR Catalogs: Insights from the Transparency Case

Roxana L. Q. Portugal[1,2][0000-0001-7693-5353], Lyrene F. Silva[3][0000-0003-1772-6062], Henrique P. S. Sousa[5][0000-0003-2150-8113] and Julio C. S. P. Leite[4][0000-0002-0355-0265]

[1] Ludwig Maximilian University of Munich (LMU), Munich, Germany
[2] Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), Cuzco, Peru
[3] Universidade Federal do Rio Grande do Norte (UFRN), Natal-RN, Brazil
[4]Universidade Federal do Rio de Janeiro (UNIRIO), Rio de Janeiro-RJ, Brazil
[5]Universidade Federal da Bahia (UFBA), Bahia-BA, Brazil
roxana.portugal@ifkw.lmu.de, lyrene.silva@ufrn.br, hsousa@uniriotec.br, julioleite@ufba.br

Abstract. The Softgoal Interdependency Graph (SIG) models non-functional requirements (NFRs) by representing softgoals and their interrelationships. However, building a SIG is challenging as it requires a deep understanding of qualitative concepts that vary across domains. The Transparency SIG (TSIG), which integrates over 30 related qualities, exemplifies this complexity. This study explores whether Large Language Models (LLMs), specifically ChatGPT-3.5 and ChatGPT-4o, can augment the knowledge of the TSIG. Through interactive dialogues, we analyzed the models' ability to suggest relevant content and structure. Our findings show that, using the TSIG as the Gold Standard, the ChatGPT-generated models demonstrated the ability to approximate the expert knowledge represented in the TSIG, as evidenced by three authors achieving over 84% recall. Furthermore, since precision varied significantly—from 29.4% to 100%—this highlights differences in the amount of false positives. These elements require further qualitative evaluation to determine which of them may actually contribute to augmenting the knowledge on transparency, as modeled by the TSIG.

Keywords: Large Language Models (LLMs), Non-Functional Requirements (NFR), Softgoals Interdependency Graph (SIG), Transparency, ChatGPT.

1 Introduction

In requirements engineering, defining quality requirements such as security, usability, and transparency is crucial, yet it remains a persistent challenge. These requirements are typically represented as softgoals in the Softgoal Interdependency Graph (SIGs), and they are inherently qualitative, abstract, and interconnected with other quality requirements. This complexity makes their elicitation, representation, and validation particularly difficult.

The elicitation of such requirements often relies on expert knowledge, domain-specific guidelines, and insights from the literature. As software systems grow more complex and interdisciplinary, there is an increasing need to analyse new softgoals and relationships, or

Proceedings of the 28th Workshop on Requirements Engineering (WER2025), August 20-22, 2025, Rio de Janeiro-RJ, Brazil.
https://doi.org/10.29327/1588952.28-11

uncover previously unrecognized links among existing ones. Consequently, it becomes essential, though challenging, to assess and refine pre-defined SIGs, with the potential to evolve them as new knowledge emerges.

In particular, the TSIG [2] results from an extensive study involving domain experts. It is defined through 33 interrelated softgoals, forming a cohesive and intricate network. The TSIG has been widely applied in academia and instantiated for various domains [5]-[13]. One key task for requirements engineers using the TSIG is to evaluate whether it adequately meets the needs and expectations of the specific context in which it is applied. However, the complex nature of transparency, intersecting with different qualities across multiple domains, presents the challenge of finding experts with comprehensive knowledge or reconciling divergent perspectives. As such, assessing the capability of LLMs (in this case, GPT) is of interest to understand their role in evolving previous results made by humans.

Our study investigates the potential of LLMs to augment knowledge about software transparency [2], as modeled through SIGs. This investigation was conducted through interactive dialogues with the ChatGPT interface, using both ChatGPT-3.5 (in 2024) and ChatGPT-4o (in 2025) [34]. According to its self-description, "I've been trained on a diverse dataset comprised of various text sources, including books, articles, websites, and other texts, to develop a broad understanding of human language and knowledge." This extensive training enables GPT to synthesize insights across domains, making it a promising tool for exploring and potentially evolving previously encoded knowledge, such as that captured in the TSIG.

The results show that the GPT-generated models can approximate the TSIG, while also revealing limitations identified through qualitative evaluation. Furthermore, the findings suggest opportunities to enrich the TSIG with new softgoals in future studies

This paper is structured as follows. Section 2 summarizes the background and presents related work. Section 3 describes the methodology applied. Section 4 presents the qualitative and quantitative results of our assessment. Section 5 discusses the outcomes of the dialogue conducted with ChatGPT and reflects on the insights gained from using GPT, understood as a broader class of LLMs. Finally, we conclude with a summary and discuss potential future avenues of research.

2 Background and Related Works

In requirements engineering (RE), qualitative concepts are defined as softgoals [4], which may be represented with SIG models of the NFR Framework [4]. The SIG helps requirement engineers model qualitative concepts and enrich knowledge by mapping softgoals that may interact or interfere with one another.

Softgoals are not only conceptually "soft," but their relationships are also challenging to represent. Consequently, the predominant relationship semantics in SIGs is help/hurt, which captures the qualitative nature of contributions. Some relationships are modeled as make/break to reflect stronger causal links between softgoals. However, the nature of these links may depend on perspective or operationalization.

Transparency is one such complex and abstract concept. It is usually defined in terms of other qualities. For example, ethical transparency is often described as clear and honest communication, linking it to principles such as integrity and trustworthiness [1]. Others define it in terms of openness and accessibility of decision-making processes, correlating transparency to the principles of accountability and fairness [2]. As such, transparency is

a quality often adapted to different domains and stakeholder interests, such as governance and public participation [3].

Thus, transparency is not a self-contained concept. It derives meaning through its relation to other qualities. The TSIG [2] (Fig. 1) reflects this interdependence, modeling transparency as a composition of 33 interconnected qualities [2]. The TSIG has been widely applied in academia and instantiated in multiple domains [5]-[13], requiring consideration of the contextual nuances of each domain.

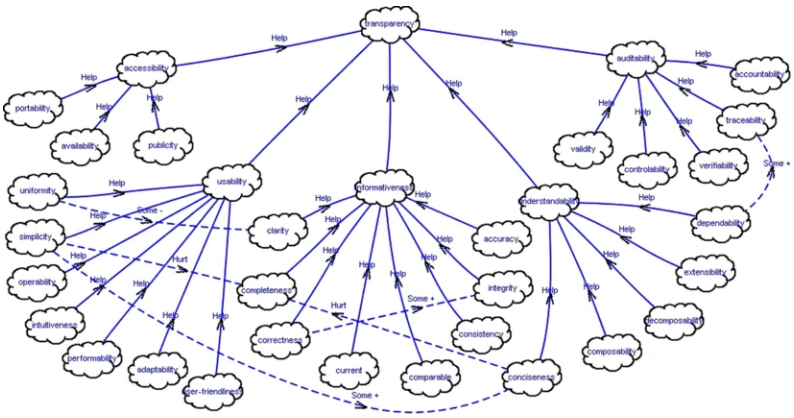

Fig. 1. Transparency Catalog [2].

To the best of our knowledge, few works specifically address the challenge of evolving SIGs [14]. Cysneiros et al. [15] proposed a frame-based representation with closeness criteria to query and identify similar softgoal interdependencies. Yamamoto [16] explored weighted softgoals to handle trade-offs by assessing their relative importance. To improve the clarity of interdependent relationships, Cleland-Huang et al. [17] introduced a goal-centric traceability method using probabilistic networks to link functional changes to SIG elements, improving impact analysis. Supakkul & Chung [18] extended the Problem Frames approach to incorporate stakeholder concerns using "soft-problems," refining and tracing them within a Problem Interdependency Graph.

With the advent of LLMs, new avenues have emerged in RE research. Several studies have explored ChatGPT's potential in eliciting requirements. Ronanki et al. [19] report that while ChatGPT outperforms experts regarding transparency and explainability, it still exhibits superficial knowledge. Zhang et al. [20] find that ChatGPT outperforms traditional information retrieval (IR) systems, as expected, given its strong semantic capabilities. In addition, Khojah et al. [21] conducted an observational study showing how engineers interact with ChatGPT, distinguishing between "prompts" and "queries" and offering insights from data ratings and exit surveys.

Building on these lines of research, Chen et al. [22] examined GPT-4's capabilities in modeling with the Goal-oriented Requirement Language (GRL). Their study addressed three research questions: how much goal modeling knowledge GPT-4 retains, how it performs in generating models from textual descriptions of varying levels of detail, and how interactive feedback influences the quality of generated models. Their evaluation was based on metrics such as 0–5 grading scales, averages, and comparisons with a "ground truth" model. Their findings suggest that GPT-4 exhibits considerable modeling knowledge, benefits from clear prompts, and

can improve with feedback. However, it also produces sometimes inaccurate or too generic responses, and its performance varies across executions.

Our own study addresses Chen et al. [22] concerns, but has a focus on the TSIG, and uses a mixed-methods approach. One of us applied precision and recall measures to compare four GPT-generated SIGs against a gold-standard reference model, reflecting a similar quantitative perspective to that of Chen et al. In addition, we carried out a qualitative analysis based on interviews with domain experts. This allowed us to identify recurring issues, such as model inconsistency, misclassification of qualities vs. operationalizations, and conceptual blending. While our results confirm several of Chen et al.'s observations, our inclusion of expert-driven viewpoint resolution [23] added deeper contextual insights into how LLMs behave when modeling complex and interrelated requirements.

3 Method

The procedure adopted involved four participants (co-authors) interacting with the LLM through questions intended to assess the LLM knowledge concerning transparency, taking the TSIG as an anchor. This approach resembles interviews, with the LLM acting as an automated 'respondent', providing answers based on the data and relationships of GPT language models and the input provided by the interviewer.

Each participant was allowed to ask approximately 15 questions in their session and was invited to request a TSIG suggestion at the end of their interaction. The participants worked independently, and after completing their sessions, they shared the files containing their conversations with ChatGPT, along with the SIGs automatically generated from their conversations, also using ChatGPT. Subsequently, they answered a questionnaire about their experience, and the results were discussed and compared (the dialogues are open access in Zenodo [34]).

The study was conducted over two years, allowing us to compare both the ChatGPT-3.5 and ChatGPT-4o models and assess their evolution in handling the TSIG. The TSIG is available in both English [2] and Portuguese [25][35], so interactions using the ChatGPT-3.5 model were conducted in both languages to explore any potential language-related differences.

This section outlines the participants' profile and evaluation criteria.

3.1 Participants profile

The four co-authors who served as participants were all experienced requirements engineering researchers, with P1 and P2 being senior researchers with over 15 years of experience, and P3 and P4 being mid-level researchers with over 10 years of experience. While all participants were familiar with the NFR framework, their familiarity with ChatGPT, SIGs, and Transparency varied, as reflected in their self-reported ratings. Participants were asked to interact with ChatGPT using approximately 15 prompts aimed at evaluating the LLM’s knowledge regarding transparency. At the end of the session, they were requested to generate a complete SIG based on the accumulated insights. Each participant answered a survey with close-ended questions about their expertise and experience, and an open-ended question regarding their general evaluation of the interaction with GPT (Section 3.2). The responses were anonymized and reviewed only after all participants had completed the survey.

For clarity, we refer to the participants as P1, P2, P3, and P4. In 2024 and 2025, participants were asked to rate their experience with ChatGPT on a scale from 1 to 5. While their expertise

in SIGs and Transparency remained unchanged across both years, their experience with ChatGPT showed slight variations. In 2024, P1 and P3 rated their experience with ChatGPT at level 4, while P2 and P4 rated it at level 3. In 2025, P2 reported an increased familiarity, rating their experience at level 4, while the other participants maintained their previous ratings. In both years, P2 and P3 indicated level 5 expertise in SIGs and Transparency, whereas P1 and P4 rated themselves at level 4 for SIGs and level 3 for Transparency.

3.2 Assessment Criteria

The results were assessed in three ways: (1) identifying participants' impressions and lessons learned, (2) focusing on identifying the characteristics of the questions posed by participants, and (3) analyzing the improvement suggestions for the TSIG provided by GPT.

For the first analysis, a survey was used to collect participants' impressions after interacting with ChatGPT. It included five closed-ended questions (using a Likert scale) and one open-ended question. The close-ended questions focused on the helpfulness of ChatGPT's responses, the model's ability to understand participants' questions, the level of trust participants had in its answers, the overall helpfulness of the interaction, and the potential for using ChatGPT in future research. The open-ended question asked participants to summarize their experience, emphasizing consistency and key learnings.

The second analysis focused on identifying the types of prompts used and the main subject of the questions. Zero-shot prompting [26] was employed to pose direct questions without prior examples, assessing the model's ability to respond about transparency without additional context. Additionally, few-shot prompting [26] was applied by providing examples and contextual information

The questions targeted 3 levels of detail in the TSIG, as summarized in Table 1:

Level 1 – Transparency-Related Qualities (softgoals) in TSIG: participants ask general questions about the concept of transparency and its core attributes.

Level 2 – Relationships among Qualities (NFRs): Participants were asked about the interdependencies between transparency-related softgoals in the TSIG.

Level 3 – TSIG Evolution: Participants asked ChatGPT about additional qualities or relationships not yet represented in the TSIG.

Note that participants were not previously instructed on the three levels of detail of the questions. However, the authors, who also served as participants, have extensive experience in SIGs and were already knowledgeable about these levels. The categorization of questions into levels was coded post-hoc during the analysis phase to classify the types of prompts posed.

In the third analysis, we compared the results of the dialogues conducted by the participants in both years. The questions were the same in both years. Additionally, participants asked ChatGPT to generate a SIG based on the information they had collected. We then evaluated how closely each generated SIGs aligned with the Gold Standard (TSIG) using recall and precision metrics. The questionnaire was designed using the work of Ronaki et al. [21] as an anchor.

	P1	**P2**	**P3**	**P4**
Prompting strategy	few-shot	zero-shot	few-shot	few-shot
Level of detail	1,2,3	2,3	1,3	1,2,3

Table 1. Question Strategy and Level of Questions used by each Participant.

It is important to note that prompting strategies were not deliberately balanced across participants. Instead, the study aimed to capture the participants' natural interaction styles with the LLM, allowing us to observe organic differences between few-shot and zero-shot approaches without introducing artificial symmetry.

4 Results

The results are derived from: (1) the classification of questions asked during participants' interactions with ChatGPT, (2) a summary of their impressions, lessons learned, and feedback gathered from the survey, (3) the SIGs generated from each participant's interactions in 2024 and 2025 years, and the evaluation of recall and precision against the Gold Standard.

4.1 Analysis of Participants' Questions (Prompts)

Based on the analysis of these interviews, we organized the responses according to the three levels of questions defined in our strategy. This approach allowed us to develop a coherent structure and perform a precise analysis of the areas of interest expressed by the participants.

As shown in Fig. 2, we categorized the 58 questions into distinct groups. The group focused on *TSIG completeness* (direct assessment of GPT knowledge concerning the possibility for the evolution of the gold standard) as the most frequently addressed by participants, with 28 questions, highlighting a strong interest in exploring missing attributes within the current SIG. The second most common category involved identifying additional *qualities related to transparency in the TSIG*, with 20 questions. Lastly, only 14 questions explored the interdependencies or *relationships among qualities* in the TSIG, suggesting that participants paid less attention to established relationships.

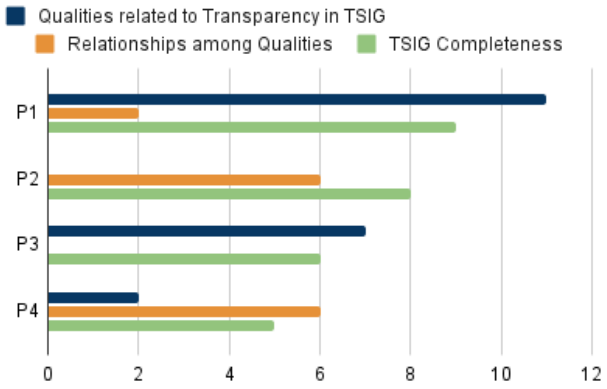

Fig. 2. Distribution of Questions by Category in the Dialogues.

4.2. Results from Questionnaire

Table 2 presents the participants' (co-authors) answers for the quantitative (closed-ended) questions of the survey. The responses remained relatively consistent across both years (Table 2). The average scores per question in 2024 were 3.75, 3.75, 2.75, 4, and 4.5, while in 2025, they showed a slight increase in the first four questions, reaching 4, 4, 3.25, and 4.25, and a minor

decrease in the last one, dropping to 4.25. This modest improvement suggests that the experience with GPT-4o did not differ significantly from that with GPT-3.5 for the participants.

It is also observed that P2 and P4 tended to be more critical in their evaluations, whereas P1 and P3 showed a more positive perception. Trust in ChatGPT's responses remained the weakest dimension, even when using the newer model.

Table 2. Quantitative Questions and Responses.

Question	2024				2025			
	P1	**P2**	**P3**	**P4**	**P1**	**P2**	**P3**	**P4**
Did ChatGPT provide helpful responses to your questions?	4	3	4	4	5	3	5	3
How did you perceive the ChatGPT level of understandability for your questioning?	3	4	4	4	5	4	4	3
What level of Trust would you assign to ChatGPT based on its answers to questions?	3	2	3	3	3	3	4	3
How helpful was the interaction with ChatGPT?	4	3	5	4	5	4	5	3
After using ChatGPT for this research, what potential do you see in using it for further research?	5	5	4	4	5	5	5	2

Regarding the open-ended questions in the survey, all participants identified inconsistencies in ChatGPT´s responses throughout their conversations [34]. This includes superficial or overly general answers and instances where the model conveyed a different context than the one intended, for example, interpreting *accessibility* and *usability* solely from the user perspective without considering their role in the broader concept of *transparency*. Additionally, responses tended to follow a standard medium length, regardless of the complexity of the question, often resulting in shallow and generic responses.

Participants also noted that ChatGPT frequently aligned itself with the framing of the question, displaying a complacent attitude that lacked critical reflection. As a result, providing too much context tended to bias the responses. Several issues were identified concerning the meaning of some softgoals, confusion between levels of granularity (e.g., mixing operationalizations with softgoals), and inaccuracies in differentiating contributions and correlations.

Many of the qualities suggested by GPT were already present in the TSIG, which gives the impression that they are validated by it; however, the dialogues show that some of these qualities were approached from a different perspective than the one used in the construction of the TSIG, a process in which at least two of this paper's authors of were involved. All participants reported that analyzing the responses provided by ChatGPT required considerable time and attention. Although the answers initially seemed useful, some of them were found to be inaccurate upon closer examination. To address these inconsistencies, participants cross-checked the LLM's suggestions against their expert knowledge and engaged in collaborative discussions after completing their individual sessions.

At the same time, the participants noted that the fact that all answers were justified made it easier to assess which parts could be taken into consideration. Despite inaccuracies and inconsistencies, some participants felt that ChatGPT provided, in some cases, useful responses regarding transparency, softgoals, and correlations. The answers were clear and explanatory,

which can be especially helpful for those unfamiliar with the concept of transparency. Additionally, participants highlighted benefits such as low cost, the ability to accelerate exploratory phases, and the potential to make complex knowledge more accessible.

4.3. Results from GPT Suggestions

Table 3 presents the SIGs created in 2025 using the GPT-4o model. The following legend compares them against the Gold Standard: (+: new; *: new position; #: merged; @: redefined). Notably, participant P3 produced the SIG with the highest number of qualities, even in the 2024 study. Due to the reduced space in the article, we have made the comparative table for the year 2024 only available on Zenodo [34].

To assess how closely the SIGs generated with ChatGPT align with the Gold Standard (TSIG), we conducted a quantitative evaluation using recall, precision, and F1-score metrics. Table 4 presents the results for all participants across both years (2024 and 2025), comparing the content of each SIG by softgoal category. For each participant (P1–P4), we show values for true positives (TP), false positives (FP), false negatives (FN), and the derived percentages of precision, recall, and F1-score.

The results show apparent differences in the coverage and precision of the SIGs generated across participants and models (Table 3).

P4 (2025) achieved the best overall performance, with an F1-score of 98.5%, closely matching the Gold Standard across all categories. In addition, it reached perfect precision (100%) for each softgoal, with only one false negative. Paradoxically, this high performance may indicate that the dialogue with ChatGPT yielded fewer novel elements to contribute to the evolution of the TSIG.

P3 (both years) generated the most extensive SIGs in terms of quantity (81 and 54 softgoals, respectively) but with low precision (35% and 52%, respectively) due to a high number of false positives. Nevertheless, recall remained high (87% - 84% respectively), indicating that the model could retrieve most of the relevant softgoals from the TSIG. The low precision reflects the inclusion of many potentially irrelevant softgoals; however, this broader output may contain valuable suggestions for augmenting the TSIG with novel qualities. Although P3 did not provide explicit structural examples, the prompting included thematic guidance, placing it near a zero-shot strategy, but ultimately closer to light few-shot prompting. This hybrid approach may have contributed to generating more diverse content, albeit with less alignment to the TSIG structure.

P1 maintained a balanced performance in both years, with F1-scores above 79% in 2025 and 81% in 2024, showing consistent understanding and prompting strategy over time. As with other participants, a relatively low precision score may indicate a broad set of potential contributions to the evolution of the TSIG.

P2 showed the lowest scores, particularly in 2025, with an overall F1-score of only 20%, likely due to a high number of omissions (FN = 28). The 2024 performance was slightly better, with an F1-score of 25%, although it was still below the average. This could be attributed to the participant's approach (zero-shot prompting), as P2 was the only one to adopt this strategy. By avoiding example-based guidance, this method may have reduced the model's bias toward reproducing the existing TSIG structure, which was more evident in the outputs of other participants. It is important to note that some softgoals may appear similar during the quantitative assessment but represent distinct concepts. For example, *timeliness* is not equivalent to *current*; *compositivity* may overlap with *composability*; *performance* and *performability* refer to different system properties; and *decomposability* should not be confused with *divisiveness*.

Table 3. Contrast among TSIG and participants' results (legend: + new; * new position; # merge; @ redefined) 2025

Gold Standard	P1	P2	P3	P4
Transparency ├── Accessibility │ ├─ Portability │ ├─ Availability │ └─ Publicity ├── Usability │ ├─ Uniformity │ ├─ User Friendliness │ ├─ Simplicity │ ├─ Operability │ ├─ Intuitiveness │ ├─ Adaptability │ └─ Performability ├── Informativeness │ ├─ Clarity │ ├─ Consistency │ ├─ Integrity │ ├─ Correctness │ ├─ Accuracy │ ├─ Current │ ├─ Completeness │ └─ Comparable ├── Understandability │ ├─ Dependability │ ├─ Composability │ ├─ Decomposability │ ├─ Extensibility │ └─ Conciseness └── Auditability ├─ Validity ├─ Controllability ├─ Verifiability ├─ Traceability └─ Accountability	Transparency ├── Accessibility │ ├─ Portability │ ├─ Availability │ └─ Publicity ├── Usability │ ├─ Uniformity │ ├─ User Friendliness │ ├─ Simplicity │ ├─ Operability │ ├─ Intuitiveness │ ├─ Adaptability │ └─ Performability ├── Informativeness │ ├─ Clarity │ ├─ Integrity │ . └─ Consistency * │ ├─ Correctness │ └─ Accuracy * │ ├─ Current │ ├─ Completeness │ ├─ Comparable │ ├─ Relevance + │ └─ Actionability + ├── Understandability │ ├─ Modularity # │ ├─ Extensibility │ ├─ Conciseness │ ├─ Explainability + │ ├─ Readability + │ └─ Visualization + └── Auditability ├─ Compliance @ ├─ Controllability └─ Accountability * ├─ Verifiability ├─ Traceability ├─ NonRepudiation + ├─ Reproducibility + ├─ Documentation + └─ Dependability *	Transparency ├── Accessibility │ ├─ Availability │ ├─ Usability * │ ├─ Inclusivity + │ └─ Responsiveness + ├── Openness + │ ├─ Relevance + │ ├─ Clarity * │ ├─ Accuracy * │ └─ Consistency * ├── Understandability │ ├─ Simplicity * │ ├─ Logical Structure + │ ├─ Visual Aids + │ └─ Contextualization + └── Accountability & Auditability @ ├─ Traceability ├─ Auditability * ├─ Security + ├─ Compliance + └─ Documentation +	Transparency ├── Accessibility │ ├─ Portability │ ├─ Availability │ ├─ Publicity │ ├─ Interoperability + │ ├─ Inclusivity + │ ├─ Multimodality + │ └─ Standardization + ├── Usability │ ├─ Uniformity │ ├─ User-Friendliness │ ├─ Simplicity │ ├─ Operability │ ├─ Intuitiveness │ ├─ Adaptability │ ├─ Performance @ │ ├─ Learnability + │ ├─ Feedback Mechanisms+ │ ├─ Error Tolerance + │ ├─ Customizability + │ ├─ Efficiency + ├── Informativeness │ ├─ Clarity │ ├─ Consistency │ ├─ Integrity │ ├─ Correctness │ ├─ Accuracy │ ├─ Timeliness @ │ ├─ Completeness │ ├─ Comparability @ │ ├─ Relevance + │ ├─ Precision + │ └─ Contextualization + ├── Understandability │ ├─ Dependency + │ ├─ Compositivity @ │ ├─ Divisibility @ │ ├─ Detail + │ ├─ Conciseness │ ├─ Clarity * │ ├─ Contextualization * │ ├─ Interpretability + │ ├─ Coherence + │ └─ Visualization + ├── Auditability │ ├─ Validity │ ├─ Controllability │ ├─ Verifiability │ ├─ Traceability │ ├─ Explanation + │ ├─ Accountability │ ├─ Documentability + │ ├─ Standardization + │ └─ Repeatability +	Transparency ├── Accessibility │ ├─ Portability │ ├─ Availability │ └─ Publicity ├── Usability │ ├─ Uniformity │ ├─ Simplicity │ ├─ Operability │ ├─ Intuitiveness │ ├─ Performance + │ ├─ Adaptability │ └─ Ease of Use @ ├── Informativeness │ ├─ Clarity │ ├─ Completeness │ ├─ Correctness │ ├─ Currency │ ├─ Comparability @ │ ├─ Consistency │ ├─ Integrity │ └─ Accuracy ├── Understandability │ ├─ Conciseness │ ├─ Composability │ ├─ Decomposability │ └─ Dependability └── Auditability ├─ Validity ├─ Controllability ├─ Verifiability ├─ Traceability └─ Accountability

Table 4. Quantitative Assessment of SIGs Retrieved Using Precision, Recall, and F1-Score Values

			2024							2025						
	Softgoal Category	Gold	Eval P1-2024	TP	FP	FN	Precision %	Recall %	F1-score %	Eval P1-2025	TP	FP	FN	Precision %	Recall %	F1-score %
P1	Transparency	5	5	5	0	0	100	100	100	5	5	0	0	100	100	100
	Accessibility	3	5	2	3	1	40	66,7	50	3	3	0	0	100	100	100
	Usability	7	8	6	2	1	75	85,7	80	7	7	0	0	100	100	100
	Auditability	5	7	5	2	0	71,4	100	83,3	9	4	5	1	44,4	80	57,1
	Understandability	5	9	5	4	0	55,6	100	71,4	6	2	4	3	33,3	40	36,4
	Informativeness	8	9	8	1	0	88,9	100	94,1	10	8	2	0	80	100	88,9
	Overall	33	43	31	12	2	72,1	93,9	81,6	40	29	11	4	72,5	87,88	79,5
	Softgoal Category	Gold	Eval P2-2024	TP	FP	FN	Precision %	Recall %	F1-score %	Eval P2-2025	TP	FP	FN	Precision %	Recall %	F1-score %
P2	Transparency	5	8	5	3	0	62,5	100	76,9	4	3	1	2	75	60	66,7
	Accessibility	3	3	0	3	3	0,0	0	—	4	1	3	2	25	33,3	28,6
	Usability	7	3	0	3	7	0,0	0	—	0	0	0	7	—	0	—
	Auditability	5	3	0	3	5	0,0	0	—	5	1	4	4	20,0	20	20,0
	Understandability	5	3	0	3	5	0,0	0	—	4	0	4	5	0,0	0	—
	Informativeness	8	3	2	1	6	66,7	25	36,4	0	0	0	8	—	0	—
	Overall	33	23	7	16	26	30,4	21,2	25,0	17	5	12	28	29,4	15,2	20,0
	Softgoal Category	Gold	SIG P3-2024	TP	FP	FN	Precision %	Recall %	F1-score %	SIG P3-2025	TP	FP	FN	Precision %	Recall %	F1-score %
P3	Transparency	5	5	5	0	0	100,0	100	100,0	5	5	0	0	100	100	100,0
	Accessibility	3	13	3	10	0	23,1	100	37,5	7	3	4	0	42,9	100	60,0
	Usability	7	16	6	10	1	37,5	85,7	52,2	12	6	6	1	50	85,7	63,2
	Auditability	5	14	5	9	0	35,7	100	52,6	9	5	4	0	55,6	100	71,4
	Understandability	5	15	3	12	2	20,0	60	30,0	10	2	8	3	20,0	40	26,7
	Informativeness	8	18	7	11	1	38,9	87,5	53,8	11	7	4	1	63,6	87,5	73,7
	Overall	33	81	29	52	4	35,8	87,9	50,9	54	28	26	5	51,9	84,8	64,4
	Softgoal Category	Gold	Eval P4-2024	TP	FP	FN	Precision %	Recall %	F1-score %	Eval P4-2025	TP	FP	FN	Precision %	Recall %	F1-score %
P4	Transparency	5	5	5	0	0	100,0	100	100,0	5	5	0	0	100	100	100,0
	Accessibility	3	0	0	0	3	—	0	—	3	3	0	0	100	100	100,0
	Usability	7	7	7	0	0	100,0	100	100,0	7	7	0	0	100	100	100,0
	Auditability	5	5	5	0	0	100,0	100	100,0	5	5	0	0	100	100	100,0
	Understandability	5	7	5	2	0	71,4	100	83,3	4	4	0	1	100	80	88,9
	Informativeness	8	8	8	0	0	100,0	100	100,0	8	8	0	0	100	100	100,0
	Overall	33	32	30	2	3	93,8	90,9	92,3	32	32	0	1	100,0	97,0	98,5

The results presented in this section highlight the value of combining both types of assessment. Quantitative metrics allow us to evaluate the degree of alignment against a defined reference, such as our Gold Standard. In many contexts, such comparison is not feasible, as there is no baseline model to contrast with what is obtained from ChatGPT. In contrast, qualitative analysis reveals the exploratory and creative potential of each interaction, uncovering valuable contributions that could enrich and expand the TSIG.

5. Discussion

Two of the authors conducted a critical review of the new elements proposed by ChatGPT in each softgoal category (Table 3). The aim was not to validate a new version of the TSIG but to examine the variety and subjectivity of the retrieved elements, thereby highlighting the challenges in constructing and validating SIGs.

Several issues were identified across categories. In *accessibility*, elements such as *inclusivity*, *responsiveness*, and *customization* were considered unrelated to the concept of *accessibility* in

the context of *transparency*. These terms align more with *usability*, whereas *accessibility* from a transparency perspective refers to the ability to reach or access the object, not necessarily to use it. *Security*, though relevant, was identified as orthogonal to *transparency*, meaning it is correlated but not directly contribute.

In *usability*, suggestions such as *error handling* and *feedback mechanisms* were regarded more as operationalizations than softgoals.

In *informativeness*, terms like *relevance* and *actionability* generated debate. While one author argued that *relevance* does not necessarily improve the ability to inform, another suggested it may help ensure that the reader receives pertinent information. Regarding *actionability*, it was pointed out that being able to act on information does not necessarily imply that one has been adequately informed. Similarly, *context sensitivity* may overlap with aspects of *usability* or *understandability*.

The category of *understandability* presented the most divergence. Terms such as *simplicity*, *clarity*, and *contextualization* were questioned in terms of their contribution to how the object of *transparency* is presented. Other terms, like *modularity*, were considered already represented by *composability* and *decomposability*. Several additions, such as *visual aids*, *logical structure*, and *explainability* were classified as operationalizations or more appropriately placed under other softgoals.

In *auditability*, elements such as *logging*, *documentation*, and *anomaly detection* were also regarded as operationalizations. Security was seen as orthogonal, while *compliance* and *dependability* sparked discussion as to whether they truly contribute to *auditability* or are instead qualities that can be assessed through it.

This review underscores the importance of distinguishing between actual softgoals and the mechanisms used to operationalize them. It also illustrates how interpreting transparency-related qualities can be highly subjective and context-dependent. Our study highlights the diversity and complexity of interpretations that arise when using an LLM to support the construction of SIGs, emphasizing the challenges of modeling nuanced concepts. The understanding of the knowledge embodied in the TSIG provided a solid foundation for critically assessing the extent of GPT's capabilities in addressing such complexity, particularly when modeling with the NFR framework (SIG).

While generating new knowledge through collaborative human effort remains a non-trivial endeavor, these limitations do not imply that LLMs are a barrier. On the contrary, they may serve as valuable catalysts for accelerating the early stages of exploration and expanding the scope of SIG construction. However, by producing a broader range of possibilities, LLMs may also increase the effort required in subsequent stages of human collaboration, particularly in reaching consensus, since more alternatives must be critically assessed and refined.

We acknowledge that the small number of participants in this study limits the generalizability of the findings. However, it is important to stress that the four participants/co-authors have a unique combination of expertise as detailed in Section 3.1. Knowledge on the topic of transparency is rare, and the qualitative reasoning explained in this Section shows how difficult it is to balance the knowledge of the representation language and of the domain we are dealing with. It is also a case that we did not conduct a traditional experiment, where the number of participants is an important issue, due to the lack of knowledge on both the representation language and the concept of transparency.

6 Conclusions

The literature on Requirements Engineering highlights NFR catalogs as an approach to managing the complexity of analyzing, documenting, and reusing Quality Requirements. However, reviewing these catalogs demands the involvement of multiple experts, significant cognitive effort to reconcile different perspectives and interpretations, and the need to ensure consistency in interdependencies [26 -29].

This study investigated the potential of using an LLM as a tool to support requirements engineers in the process of creating a SIG for a given quality (NFR), specifically focusing on reviewing a SIG on software transparency. To the best of our knowledge, this is the first work to tackle this approach. The choice to focus on transparency relies on its central role in trustworthy AI [8, 11, 30, 31].

The contributions of this paper to Requirements Engineering are twofold: (1) a report on the strategies employed by participants to extract new information for the graph, along with the corresponding outcomes, and (2) a quantitative and qualitative evaluation of participants' experiences using GPT. All four participants in the study acknowledged GPT's utility and potential as a valuable information source in the requirements engineering process, particularly during exploratory phases, and can, with caution, provide valuable insights for evolving NFR catalogs. We also recognize that the open-access nature of our study facilitates reproducibility [34], especially regarding the prompting process, thus encouraging other researchers to explore LLMs for evolving NFR catalogues.

In the future, we plan to focus on customizing the Transparency SIG according to different domains and exploring the NFR topic of the NFR Framework [4] since the work presented here is centered on the NFR type. In doing this, we could better use the broader LLM knowledge in different domains.

On the other hand, we will explore LLMs' capability in detecting possible operationalizations of a SIG. In this work, we explored the upper level of the TSIG (Fig. 1), whereas the software catalog [25] already has a set of these operationalizations. Of course, these operationalizations are closely related to the domain (topic) to which the NFR is being considered.

We plan to improve our three-level conversational strategy presented in this article by trying different patterns as we continue to explore the world of LLMs to help the requirements engineer use LLMs to analyze softgoals and NFR catalogs, as well as to define processes and best practices for using such patterns. One possibility is to improve our dialogues by explaining to the LLM the rationale used in our questions, as such helping the LLM focus on the specific context, as well as providing it with a basis for better explanations.

ACKNOWLEDGMENTS

Leite thanks the support of CNPq, and Portugal thanks the Institute for Communication Studies and Media Research (IfKW) at LMU Munich for its support.

References

1. Maister, D.H., Galford, R., Green, C.: The Trusted Advisor. Free Press, New York (2021).

2. Leite, J.C.S.P., Cappelli, C.: Software transparency. Bus. Inf. Syst. Eng. 2, 127–139 (2010). doi: doi.org/10.1007/s12599-010-0102-z
3. Noveck, B.S.: Smart Citizens, Smarter State: The Technologies of Expertise and the Future of Governing. Harvard University Press, Cambridge, MA (2015). doi: doi.org/10.4159/9780674915435.
4. Chung, L., Nixon, B.A., Yu, E., Mylopoulos, J.: *Non-Functional Requirements in Software Engineering*. Springer, Berlin (2012), doi: 10.1007/978-1-4615-5269-7.
5. Portugal, R.L.Q, Engiel, P., Roque, H., Leite, J.C.S.P.: Is there a demand of software transparency? In: Proceedings of the XXXI Brazilian Symposium on Software Engineering (SBES 2017), pp. 204–213 (2017), doi: 10.1145/3131151.3131155.
6. Zinovatna, O., Cysneiros, L.M.: Reusing knowledge on delivering privacy and transparency together. In: 5th IEEE International Workshop on Requirements Patterns (RePa 2015), pp. 17–24. IEEE, New York (2015), doi: 10.1109/RePa.2015.7407733.
7. Cysneiros, L.M., Raffi, M., Leite, J.C.S.P.: Software transparency as a key requirement for self-driving cars. In: 26th IEEE International Requirements Engineering Conference (RE 2018), pp. 382–387. IEEE, New York (2018), doi: 10.1109/RE.2018.00-21.
8. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., Kujala, S.: Transparency and explainability of AI systems: from ethical guidelines to requirements. Inf. Softw. Technol. 159, 107197 (2023). 10.1016/j.infsof.2023.107197
9. Hosseini, M., Shahri, A., Phalp, K., Ali, R.: Four reference models for transparency requirements in information systems. Requir. Eng. 23, 251–275 (2018), doi: 10.1007/s00766-017-0265-y.
10. Hochstetter, J., Vairetti, C., Cares, C., García Ojeda, M., Maldonado, S.: A transparency maturity model for government software tenders. IEEE Access 9, 45668–45682 (2021), doi: 10.1109/ACCESS.2021.3067217.
11. Chazette, L., Schneider, K.: Explainability as a non-functional requirement: challenges and recommendations. Requir. Eng. 25(4), 493–514 (2020), doi: 10.1007/s00766-020-00333-1
12. Carvalho, L.P., Santoro, F., Cappelli, C.: Using a citizen language in public process models: the case study of a Brazilian university. In: Knahl, M., Klein, G., Rinderle-Ma, S. (eds.) Electronic Government and the Information Systems Perspective. EGOVIS 2016. LNCS, vol. 9820, pp. 123–134. Springer, Cham (2016), doi: 10.1007/978-3-319-44159-7_9.
13. Vössing, M., Kühl, N., Lind, M., Satzger, G.: Designing transparency for effective human-AI collaboration. Inf. Syst. Front. 24(3), 877–895 (2022), doi: 10.1007/s10796-022-10284-3
14. Arango, G., Freeman, P.: Application of artificial intelligence. ACM SIGSOFT Softw. Eng. Notes 13(1), 32–38 (1988). 10.1145/43857.43869
15. L. M. Cysneiros, K. K. Breitman, C. Lopez and H. Astudillo, "Querying Software Interdependence Graphs," *2008 32nd Annual IEEE Software Engineering Workshop*, Kassandra, Greece, 2008, pp. 108-112, doi: 10.1109/SEW.2008.28.
16. S. Yamamoto, "An Approach for Evaluating Softgoals Using Weight," in *Information and Communication Technology*, I. Khalil, E. Neuhold, A. M. Tjoa, L. D. Xu, and I. You, Eds., Cham: Springer International Publishing, 2015, pp. 203–212. doi: 10.1007/978-3-319-24315-3_20.
17. Cleland-Huang, J., Settimi, R., BenKhadra, O., Berezhanskaya, E., Christina, S.: Goal-centric traceability for managing non-functional requirements. In: Proceedings of the 27th International Conference on Software Engineering (ICSE 2005), pp. 362–371. ACM, New York (2005), doi: 10.1145/1062455.1062525.
18. Supakkul, S., Chung, L.: Extending problem frames to deal with stakeholder problems: an agent- and goal-oriented approach. In: Proceedings of the 2009 ACM Symposium on Applied Computing (SAC 2009), pp. 389–394. ACM, New York (2009), doi: 10.1145/1529282.1529366.

19. Ronanki, K., Berger, C., Horkoff, J.: Investigating ChatGPT's potential to assist in requirements elicitation processes. In: 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2023), pp. 354–361. IEEE, New York (2023), doi: 10.1109/SEAA60479.2023.00061.
20. Zhang, J., Chen, Y., Liu, C., Niu, N., Wang, Y.: Empirical evaluation of ChatGPT on requirements information retrieval under zero-shot setting. In: International Conference on Intelligent Computing and Next Generation Networks (ICNGN 2023), pp. 1–6. IEEE, New York (2023), doi: 10.2139/ssrn.4450322.
21. Khojah, R., Mohamad, M., Leitner, P., Gomes de Oliveira Neto, F.: Beyond code generation: an observational study of ChatGPT usage in software engineering practice. Proc. ACM Softw. Eng. 1(FSE), 1819–1840 (2024), doi: 10.1145/3660788.
22. Chen, B., Chen, K., Hassani, S., Yang, Y., Amyot, D., Lessard, L., Mussbacher, G., Sabetzadeh, M., Varró, D.: On the use of GPT-4 for creating goal models: an exploratory study. In: 31st IEEE International Requirements Engineering Conference Workshops (REW 2023), pp. 262–271. IEEE, New York (2023), doi: 10.1109/REW57809.2023.00052.
23. Leite, J.C., Freeman, P.A.: Requirements validation through viewpoint resolution. IEEE Trans. Softw. Eng. 17(12), 1253–1269 (1991), doi: 10.1109/32.106986
24. OpenAI: ChatGPT, version 3.5. OpenAI, San Francisco, CA. https://openai.com/chatgpt (accessed: March 2024)
25. Software Transparency Catalog: http://transparencia.inf.puc-rio.br/wiki/index.php/Catálogo_Transparência (accessed: March 2024)
26. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901 (2020). http://arxiv.org/abs/2005.14165
27. Zave, P., Jackson, M.: Four dark corners of requirements engineering. ACM Trans. Softw. Eng. Methodol. 6(1), 1–30 (1997), doi: 10.1145/237432.237434
28. Carvalho, R.M., Andrade, R.M.C., Lelli, V., Silva, E.G., de Oliveira, K.M.: What about catalogs of non-functional requirements? In: Joint Proceedings of REFSQ-2020 Workshops, Doctoral Symposium, and Research Projects Track, p. 59 (2020). CEUR-WS.org, Aachen. https://ceur-ws.org/Vol-2584/PT-paper8.pdf
29. Liu, C.-L.: Ontology-based conflict analysis method in non-functional requirements. In: 9th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2010), pp. 491–496. IEEE, New York (2010), doi: 10.1109/ICIS.2010.26.
30. Cortesi, A., Logozzo, F.: Abstract interpretation-based verification of non-functional requirements. In: Canal, C., Zavattaro, G. (eds.) Coordination Languages and Models. COORDINATION 2005. LNCS, vol. 3454, pp. 49–62. Springer, Berlin, Heidelberg (2005), doi: 10.1007/11417019_4.
31. Matsumoto, Y., Shirai, S., Ohnishi, A.: A method for verifying non-functional requirements. Procedia Comput. Sci. 112, 157–166 (2017), doi: 10.1016/j.procs.2017.08.006.
32. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: from principles to practices. ACM Comput. Surv. 55(9), 1–46 (2023). 10.1145/3555803
33. Chazette, L., Brunotte, W., Speith, T.: Explainable software systems: from requirements analysis to system evaluation. Requir. Eng. 27(4), 457–487 (2022), doi: 10.1007/s00766-022-00393-5
34. Portugal, R.L.Q., Silva, L., Sousa, H., Leite, J.C.S.P.: Data: Exploring LLMs on Supporting the Elicitation of Knowledge for NFR Catalogs: Insights from the Transparency Case. In: Workshop on Requirements Engineering 2025 (WER25), Rio de Janeiro, Brazil. Zenodo (2025). https://doi.org/10.5281/zenodo.15623200
35. Conceição, J., Leite, J.C.S.P., Pitangueira, R.: O uso de inspeção para aumentar a transparência de processos. In: Proceedings of the 26th Workshop on Requirements Engineering (WER 2023), Porto Alegre, Brazil (2023), doi: 10.29327/1298356.26-15