2.3. Criteria for assessing the quality of research questions

As part of the process of formulating research questions, the evaluation of their quality was carried out using the FINER method [7], a widely recognized approach for formulating and evaluating the quality of research questions in various fields of study. This method is based on five specific criteria: Feasibility, Interest, Novelty, Ethics, and Relevance [7], which provide a comprehensive framework to ensure that the questions formulated are appropriate, feasible, original, relevant, and ethical. For the evaluation of each criterion, a score of 1 point was assigned if the question fully met the criterion, 0.5 points if it partially met the criterion, and 0 points if it did not meet the criterion, a question was considered acceptable if its average total score was between 3. Those questions that did not meet this threshold were reviewed and, if necessary, modified to ensure their relevance and pertinence to the SLR in question:

- **Feasibility (F):** Is it possible to address this question with available resources, such as access to relevant literature and analytical capacity?
- **Interesting (I):** Does the question address a topic of interest and relevance to the academic and industry community?
- **Novel (N):** Does the question explore a unique or innovative aspect that has not been widely addressed in the existing literature?
- **Ethical (E):** Will the research arising from this question be conducted in an ethical and transparent manner, respecting the principles of academic and professional integrity?
- **Relevant (R):** Is the question relevant to the field of study and does it have significant implications for software development practice and research?

To carry out the FINER method, 5 expert evaluators in the area of Software Engineering, especially in the area of process improvement and quality were involved, who helped to obtain the level of relevance and pertinence of the research questions for the SLR. This evaluation was carried out using a survey in Google Forms, through which a set of results were obtained to later analyze the results obtained, which are summarized in Table 3.

Table 3. Quality assessment criteria for research questions.

Id	Question	E	Scores per person for each of the quality criteria					Total	P
			Feasible	Interesting	New	Ethical	Relevant		
PI1	What are the types of solutions proposed in the literature to analyze and/or manage process debt in software development?	1	1	1	0.5	0.5	1	4	3.7
		2	0.5	1	1	1	1	4.5	
		3	0.5	0.5	0	0	0.5	1.5	
		4	1	1	0.5	1	1	4.5	
		5	0.5	1	0.5	1	1	4	
PI2	What are the types of research identified in the literature?	1	1	1	1	1	1	5	4.1
		2	1	1	0.5	0.5	1	4	
		3	0.5	1	0.5	0	0.5	2.5	
		4	1	1	1	0.5	1	4.5	
		5	1	1	0.5	1	1	4.5	
PI3	What is the current level of maturity and/or development in the literature regarding process debt management concepts and approaches in software development?	1	1	1	1	1	1	5	4.5
		2	1	1	1	1	1	5	
		3	1	1	1	0	0.5	3.5	
		4	1	1	1	0.5	1	4.5	
		5	1	1	0.5	1	1	4.5	
		1	1	1	1	1	1	5	

PI4	What are the benefits, challenges, causes, impact, effects and/or consequences associated with process debt research in software development?	2	0.5	0.5	0.5	0.5	0.5	2.5	3.9
		3	1	1	0.5	0	0.5	3	
		4	1	1	1	0.5	1	4.5	
		5	1	1	0.5	1	1	4.5	

Acronyms used: E: evaluator, P: Average.

Based on the answers obtained using the FINER method, the evaluators consider that the research questions meet the quality criteria established by the method, and although it can be observed that the score of the questions did not reach the maximum expected, they managed to exceed the lower limit of the threshold previously established.

2.6. Assessing the relevance and pertinence of primary studies

Evaluating the relevance and pertinence of primary articles for research allows minimizing biases and maximizing the internal and external validity of a research [5]. Following this premise and with the purpose of maintaining scientific rigor in the research, the relevance and pertinence of the selected primary articles were evaluated based on an adaptation of the instrument proposed by Kitchenham and Charters [5] Kitchenham *et al* [8]. The assessment of relevance and pertinence was based on three main factors: (i) internal validity, which refers to the ability of the study to adequately answer the research question posed; (ii) external validity, which according to Dybá *et al*. [9], allows analyzing the generalizability of the results for the population of interest and the applicability of the findings; (iii) bias, defined by Dybá *et al*. [9] as a systematic error or deviation from the truth in the results or inferences, which reflects a tendency to generate results that are systematically far from reality. Considering the above, an instrument was designed for this evaluation that includes a checklist covering four categories: clarity, rigor, credibility, and relevance of each study. Table 5 presents the questions used to measure the level of relevance and pertinence based on the four categories established. These categories are described below:

- **Clarity**: This category emphasizes the importance of presenting SLR results in a clear and concise manner so that they can be understood by other researchers and practitioners in the field of software engineering [10].
- **Rigor:** This category emphasizes the need to follow a rigorous and reliable methodology throughout the SLR process to ensure the validity of the results obtained [6].
- **Credibility:** This category allows knowing the level at which the research results are meaningful, valid and effective by evaluating the scientific methods used, discussing the limitations of the research process and analyzing the results obtained [11].
- **Relevance:** This category allows determining whether the results of the study are valid, useful, well presented and significant. In addition, it identifies the scientific methods used and the analysis of the findings found [9].

Table 5. Relevance and pertinence of primary articles.

Id	Relevance and Relevance of Primary Items	Scoring for answers			Category
		+1	0	-1	
1	Does the article clearly focus on process debt in software development?	Yes	Partially	No	Clarity
2	Does the article clearly describe the research problem and objectives?	Yes	Partially	No	

3	Does the article provide a clear and operational definition of process debt?	Yes	Partially	No	
4	Does the article follow a formal, structured research methodology?	Yes	Partially	No	Strictness
5	Are the data collection and analysis methods appropriate and well described?	Yes	Partially	No	
6	Does the article provide sufficient detail to allow replication of the study?	Yes	Partially	No	
7	Does the article have clear practical implications for the software industry?	Yes	Partially	No	Relevance
8	Does the article contribute significantly to academic knowledge about process debt?	Yes	Partially	No	
9	Does the article propose future work or additional areas of research?	Yes	Partially	No	
10	Does the article use scientific methods appropriate to the research?	Yes	Partially	No	Credibility
11	Does the article discuss the limitations and biases of the research process?	Yes	Partially	No	
12	Are the results of the article meaningful and valid?	Yes	Partially	No	

3.1 Evaluation of relevance and pertinence

The assessment of the relevance and pertinence of the primary studies, detailed in Table 10, not only reflects the theoretical foundation of the selected articles, but also provides an in-depth understanding of the areas where process debt research can still make progress. Each study was evaluated based on twelve criteria to determine the importance of its contribution to the understanding and management of process debt. The scores obtained, ranging from -12 to 12 points, reflect both the value and possible limitations of each article, making visible the diversity of approaches in this emerging area.

Figure 1 provides a clear visual grouping of the evaluated studies, and its relevance lies in highlighting how the different approaches have been rated in terms of relevance to the academic community. It is notable that only 5% of the studies [2] achieved the maximum score of 12 points, which could indicate that there are outstanding approaches that serve as key benchmarks in process debt management. This fact underlines the need to look in detail at what were the criteria and strategies that allowed this article to stand out so significantly above the rest. On the other hand, 5% of the studies [1] achieved a score of 11 points, while 15% [3], [16], [21] achieved 10 points, suggesting that there is a relatively small group of investigations that, although they do not reach the highest score, offer important contributions to process debt management. These articles stand out as significant contributions that encourage interested readers to explore further.

The distribution of studies with intermediate scores, such as those that scored between 6 and 9 points. Some 10% of the studies [24], [19] scored 9 points, while another 10% [14], [29] scored 7 points, showing approaches that, while effective, could benefit from deeper integration with real software development contexts. Nevertheless, these studies provide a valuable foundation on which to build further, and it would be advisable to examine them in more detail to identify possible improvements or areas of application in enterprise environments.

Figure 1 also highlights the presence of articles with lower scores, such as those with 5 points or less, which represent 40% of the studies evaluated. This group includes articles with negative scores, such as the study [22] (-1) and [26] (-6), which opens a space for critical reflection on the relevance of their contributions. These studies not only indicate a lack of

alignment with the evaluated criteria, but also invite a deeper reading to understand whether the low score is due to methodological, contextual or focus limitations.

Given the difference in scores, both Table 10 and Figure 1 should be reviewed carefully, as both provide a visual representation that facilitates the identification of patterns of relevance across studies. These differences in scores should be viewed as an opportunity to assess the current state of process debt research. In this sense, this is where the academic community can focus future efforts, for example: fostering interdisciplinary collaboration and developing theories that improve current management practices [2]. Similarly, case studies could be conducted in companies, training programs could be promoted and collaboration with industry could be strengthened [3]. In addition, it is essential to develop conceptual frameworks, adopt holistic approaches that integrate different forms of debt and case studies in real environments [17].

Table 10. Evaluation of the relevance and pertinence of the primary articles.

Id	Relevance and pertinence of primary articles												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
A1	1	1	1	1	1	0	1	1	1	1	1	1	11
A2	1	1	1	1	1	1	1	1	1	1	1	1	12
A3	1	1	1	0	0	1	1	1	1	1	1	1	10
A4	0	1	0	1	1	0	0	0	1	1	0	1	6
A5	1	1	1	1	0	0	1	1	1	1	0	1	9
A6	-1	1	-1	1	1	0	1	0	1	1	0	1	5
A7	0	1	0	1	1	0	0	0	1	1	0	0	5
A8	-1	0	-1	0	-1	-1	1	-1	0	-1	-1	0	-6
A9	0	1	-1	1	0	0	0	0	1	1	1	0	4
A10	0	1	0	1	1	0	1	0	1	1	0	1	7
A11	-1	1	-1	0	0	-1	1	-1	1	0	0	0	-1
A12	1	1	1	0	0	-1	1	1	1	0	0	1	6
A13	1	1	1	1	1	0	1	1	1	1	0	1	10
A14	0	1	0	1	1	0	1	0	1	1	1	0	7
A15	1	1	1	1	0	0	1	1	1	0	0	1	8
A16	-1	1	-1	1	1	0	0	-1	1	1	0	0	2
A17	1	1	0	1	1	0	1	1	1	1	0	1	9
A18	0	1	0	1	0	0	1	0	1	0	1	0	5
A19	1	1	1	1	1	0	1	1	1	1	0	1	10
A20	0	1	0	1	0	-1	0	0	1	0	-1	0	1

Acronyms used: Id: item identification (see Table 9).

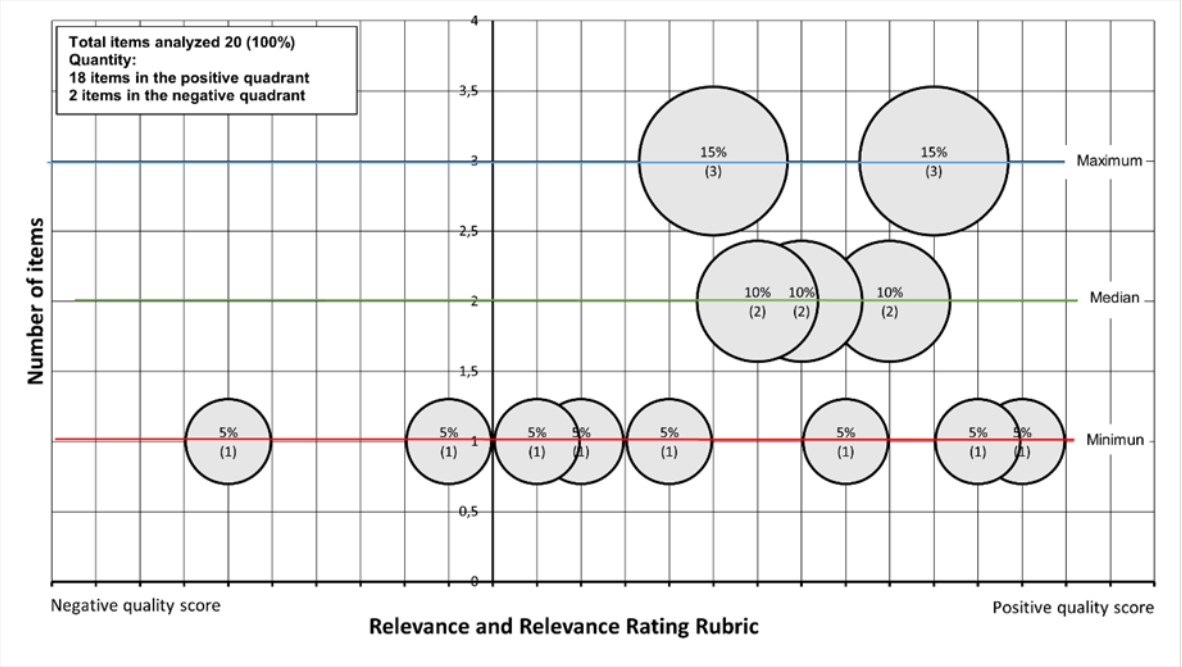

Figure 1. Visualization of the relevance and pertinence of primary articles in a bubble chart.}

5.2. Limitations of the review

- **Biases in the search string.** The initial search string used in this review was quite extensive, which resulted in the exclusion of some relevant papers. To address this limitation, it was decided to use a shorter and more general search string, which allowed us to cover a larger number of studies and obtain a more complete view of the topic. Different search engines were searched, including Google Scholar, Scopus, Science Direct, IEEE Xplore, Web of Science, ACM and Springer Link. Finally, Google Scholar was selected as the primary search engine because of its ability to comprehensively cover relevant studies that were also found in other databases. However, this focus on a single main search engine may have introduced bias by relying heavily on its algorithm.
- **Biases in the selection of studies.** The selection of studies was based on predefined inclusion and exclusion criteria, which could have led to the omission of relevant research that did not meet all the established criteria. This selection process is subject to subjective biases, where the interpretation of the criteria could have influenced the decision to include or exclude certain studies. The accessibility of the documents was also a relevant factor in the selection, as documents from undergraduate, master's and doctoral theses were included, as well as types of publications ranging from conferences, scientific articles and gray literature. This diversity in the types of documents allowed for a broader view of the topic, although it could also introduce a bias by depending on the availability and accessibility of these materials.
- **Biases in data extraction.** To minimize biases in data extraction, a structured protocol was followed to select the primary articles. This protocol included a careful evaluation of the topic covered, the introduction of each article and keywords, as well as defined inclusion and exclusion criteria. Although this systematic approach helped to ensure the relevance and quality of the selected studies, subjective interpretation of these elements may have introduced bias. It is possible that greater weight was given to certain studies based on their alignment with the objectives of the review, which could have influenced the way in which the data were extracted and presented.

- **Biases in the interpretation of results.** The interpretation of the results may have been influenced by prior expectations and by the limited number of studies available that directly address process debt. This phenomenon may have led to overinterpretation of certain findings or underestimation of others, depending on how the results matched the initial hypotheses. This limits the generalizability of the results to a broader context, which is an aspect to be taken into account in future research.