Hierarchical Evaluation Framework: Best Practices for Human Evaluation

Iva Bojic[1] and **Jessica Chen**[2] and **Si Yuan Chang**[1] and **Qi Chwen Ong**[1] and **Shafiq Joty**[1,3] and **Josip Car**[1,2]
[1]Nanyang Technological University Singapore
[2]Imperial College London, United Kingdom
[3]Salesforce Research, USA

Abstract

Human evaluation plays a crucial role in Natural Language Processing (NLP) as it assesses the quality and relevance of developed systems, thereby facilitating their enhancement. However, the absence of widely accepted human evaluation metrics in NLP hampers fair comparisons among different systems and the establishment of universal assessment standards. Through an extensive analysis of existing literature on human evaluation metrics, we identified several gaps in NLP evaluation methodologies. These gaps served as motivation for developing our own hierarchical evaluation framework. The proposed framework offers notable advantages, particularly in providing a more comprehensive representation of the NLP system's performance. We applied this framework to evaluate the developed Machine Reading Comprehension system, which was utilized within a human-AI symbiosis model. The results highlighted the associations between the quality of inputs and outputs, underscoring the necessity to evaluate both components rather than solely focusing on outputs. In future work, we will investigate the potential time-saving benefits of our proposed framework for evaluators assessing NLP systems.

1 Introduction

Human evaluation is crucial for assessing the quality, validity, and performance of Natural Language Processing (NLP) systems especially as automatic metrics are usually not sufficient (Van Der Lee et al., 2019). Human evaluation can deal with complex generated natural language and its nuances such as pragmatics, context and semantics which often requires some expert knowledge (Sudoh et al., 2021). Automatic evaluation may be used to assess individual dimensions (e.g., fluency, accuracy) of natural language, however, may often lose to humans in terms of accuracy and understanding.

Various methodologies are often employed in human evaluation such as ranking, pairwise comparison, or a state-of-the-art machine translation metric that was used in Castilho (2021). They can provide valuable insights into the strengths and limitations of an NLP system; however, it is notably time-consuming and expensive and significant trade-offs may exist in consideration of different goals or requirements (Zhang et al., 2020). The human evaluation also comes with its own set of limitations, such as fatigue effect (van der Lee et al., 2021) and inconsistencies between evaluators. The role of human evaluators should also be considered as some tasks may require domain expert knowledge or provide specific training evaluators.

There is currently a lack of consensus on which metrics to use for the human evaluation of NLP systems (Paroubek et al., 2007). As there tend to be different research goals, requirements and task-dependent metrics, there exists the challenge of standardizing human evaluation metrics and essentially reaching an overall consensus. A unique combination of metrics can be used for a more comprehensive assessment depending on the desired objectives. These combinations can be grouped based on different evaluation aspects (Liang and Li, 2021). Metrics may also vary depending on the task (e.g., machine translation, sentiment analysis) and thus task design can affect the criteria used for evaluation (Iskender et al., 2021).

To identify gaps in the literature pertaining to human evaluation, we conducted a scoping review to systematically examine various aspects of human evaluation experiments in NLP tasks, including the characteristics of evaluators, evaluation samples, scoring methods, design of evaluation and statistical analysis. The findings of our literature review revealed three significant gaps: (i) the absence of evaluation metrics for NLP system inputs, (ii) the lack of consideration for interdependencies among different characteristics of assessed NLP systems, and (iii) a limited utilization of metrics for extrinsic evaluation of NLP systems.

We hope to bridge the aforementioned gaps by providing a standardized human evaluation framework that can be used across different NLP tasks. Our proposed framework employs a hierarchical structure that divides the human evaluation process into two phases: testing and evaluation. This division enables evaluators to assess the quality of inputs used by testers when evaluating NLP systems. Furthermore, the hierarchical design of the evaluation metric allows for the computation of a composite score that reflects the overall quality of the NLP system.

This paper is organized as follows. Section 2 presents the analysis from a scoping review that included more than 200 papers published within the last three years in the top 5 NLP venues. The results of the aforementioned analysis informed the development of the proposed hierarchical evaluation framework, which is presented in Section 3. Section 4 presents the results of adopting the proposed framework for the human evaluation of the Machine Reading Comprehension (MRC) system developed as a part of the human-AI symbiosis model. Finally, Section 5 concludes the paper.

2 Scoping Review

2.1 Structured Review

To inform our development of a hierarchical framework for human evaluation, we conducted a scoping review to examine existing literature systematically. Our paper selection process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for Scoping Reviews checklist (PRISMA-ScR) (Peters et al., 2015) (see Figure 1). We searched for relevant publication venues on Google Scholar. We selected the category of Engineering and Computer Science, followed by the sub-category of Computational Linguistics. Subsequently, we chose the top five venues with the highest h5-index, namely:

- Meeting of the Association for Computational Linguistics (ACL),
- Conference on Empirical Methods in Natural Language Processing (EMNLP),
- Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL),
- Conference of the European Chapter of the Association for Computational Linguistics (EACL),
- International Conference on Computational Linguistics (COLING).

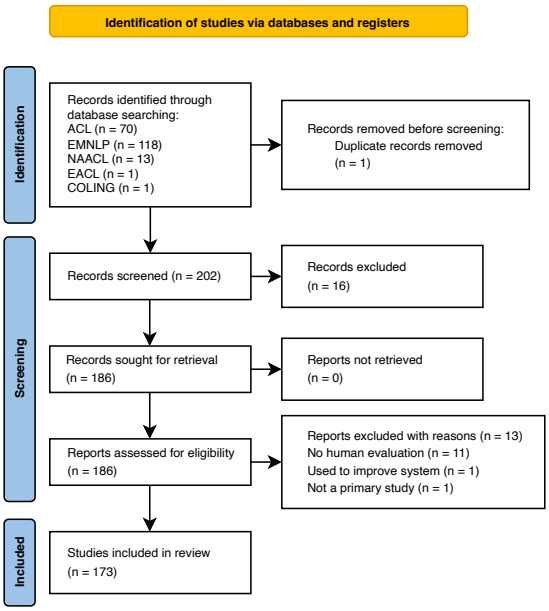

Figure 1: This PRISMA flow diagram depicts the study selection process throughout this scoping review. 203 studies in total were identified through a search on Google Scholar. After one duplicate was removed, the total remaining studies was 202. After title and abstract screening, 16 studies were excluded, leaving 186 studies for full-text screening. A final 173 studies were included in this scoping review for data extraction and analysis.

Due to the rapid development in the NLP field, only studies published between 2019 and 2023 were included. The Google Scholar search strategy is shown in Figure 2.

2.2 Selection of Articles

Eligible articles were identified in two stages: (1) title and abstract screening, (2) full-text screening. To maintain consistency of decision-making in the selection process, both title and abstract screening and full-text screening were conducted by two of the three reviewers (IB, JC, QCO) independently based on pre-defined inclusion and exclusion criteria (see Figure 3). Conflicts were resolved through discussion with a third reviewer to establish consensus. The resolution of inconsistencies or disagreements amongst reviewers was guided by predefined eligibility criteria and reference to initial objectives. Reasons for exclusion were recorded during full-text screening.

Hierarchical Human Evaluation Framework Search Strategy
(Literature Search performed: April 24, 2023)

1. "human evaluation" source:"ACL" OR source:"EMNLP" OR source:"NAACL" OR source:"EACL" OR source:"COLING"
2. "human evaluation" source:"ACL"
3. "human evaluation" source:"EMNLP"
4. "human evaluation" source:"NAACL"
5. "human evaluation" source:"EACL"
6. "human evaluation" source:"COLING"
7. Limit 1-6 to yr=2019-current

Figure 2: Search strategy used for the scoping review. After performing 1, we also performed 2-6 to find all papers from individual venues that did not appear after the first combined search.

Inclusion criteria:

1. It is a full-text article that reported empirical research in NLU, NLG or both.
2. It reported human evaluation for the purpose of evaluating the performance of the system.
3. It was published in English.
4. It was published in 2019 or later.
5. It is a peer-reviewed article published in ACL, EMNLP, NAACL, EACL or COLING.

Exclusion Criteria:

1. It reported secondary research such as a literature review, rapid review, systematic review, or scoping review.
2. It is a pre-print article, book chapter, conference abstract, expert opinions, perspectives, or commentary.
3. Human evaluation was conducted for other purposes, such as improving the system.
4. It was published in a language other than English.
5. It was published before 2019.
6. It does not involve an NLP system.
7. It was published in other venues that are not listed above.

Figure 3: This figure lists the inclusion and exclusion criteria that formed the basis of our screening process.

2.3 Data Extraction

A standardized data extraction form (see Appendix 1) was developed through iterative discussions between three reviewers (IB, JC, QCO) based on insights gained during the initial literature review of related work. The data extraction form was first piloted on three randomly selected articles by the three reviewers to ensure consistent and accurate extraction of data. The data extraction process involved all three reviewers and was done independently. Ambiguities or uncertainties were resolved by discussion between reviewers and by referring to the original papers used for the creation of the extraction matrix (Van Der Lee et al., 2019; Amidei et al., 2018a; Liang and Li, 2021; Howcroft et al., 2020). We extracted a range of variables from certain chosen sources and tailored them to the objectives of our review. These variables are categorized as follows in Section 2.4: (1) characteristics of evaluators, (2) evaluation samples, (3) scoring methods, (4) design of evaluation and (5) statistical analysis.

2.4 Synthesis of Results

2.4.1 Characteristics of Evaluators

A large proportion of papers (83%, 144/173) provided information on the number of evaluators that participated in the human evaluation. This shows that there is a general consistency in the reporting of human evaluation methods across all papers reviewed. The number of evaluators employed can be defined as *small* (1-5), *medium* (6-9) and *large* ($\geq$ 10) scale (van der Lee et al., 2021). Papers reported a small number of evaluators in 62% of cases (107/173), a medium number in 6% (11/173), and a large number in 15% (26/173). The median number of evaluators was three per study.

71% of the reviewed papers (122/173) reported the background of the evaluators, differentiating between *experts* and *non-experts*, detailed which platform they were from or set standards for crowd-sourced workers. One example, proposed in Zhu et al. (2020), was to set standards by only using workers with a high enough approval rate to ensure quality. This helps alleviate the problem of quality control when using larger-scale crowd-sourcing platforms such as Amazon Mechanical Turk.

2.4.2 Evaluation Samples

All of the papers reported that human evaluation was done only on *outputs* of NLP systems, with the median number of evaluation instances being 100. Most papers (60%, 103/173) created samples *randomly*, but some (3%, 6/173) specified *their methodology*. For instance, in Zeng and Nie (2021), discussions that were difficult to understand were filtered out. In this case, human evaluation was used to compare the dialogue generation between two different models. In order to create a more relevant dataset for human evaluation, filtering out professional texts that were difficult to understand, ensured that the data was closer to daily dialogue. This allowed for more accurate and reproducible human evaluation results. Using alternative methods to random sampling can have certain benefits such as cost-effectiveness, time efficiency and focused research objectives (Zeng and Nie, 2021).

2.4.3 Scoring Methods

Overall, 68% of papers (118/173) used a *scale* as their evaluation scoring system. A scoring system should also be defined by assigning attributes or certain qualities to a number in the scale that they are using. Further, 23% of papers (39/173) reported using *comparison* between different models or question answering to achieve more qualitative results. Examples include win, tie, loss, A/B testing, and a direct comparison.

The characteristics of evaluation can be referred to as evaluation attributes or text quality dimensions such as *fluency*, *adequacy*, and *grammar* (Gehrmann et al., 2023). These characteristics can be considered for both qualitative and quantitative methods and are often specified to guide the evaluation task. For example, Liang and Li (2021) divided various characteristics into seven groups based on their similarity and overall purpose for the human evaluation of chatbots. These groups further tailor the characteristics of evaluation to the unique task, allowing the reader to understand the reason for their selection.

Dependencies can exist among characteristics of evaluation. In other words, human evaluation can be done in sequential order when the order in which characteristics are evaluated matters. Moreover, evaluation can be prematurely stopped if some characteristics were not deemed of a satisfactory quality. Consequently, dependencies among characteristics of evaluation could also allow for a NLP system to have a composite score that would reflect its overall quality. For instance, an overall performance score can be produced based on pre-defined threshold criteria that need to be fulfilled. This threshold could be a specified performance level reached by a specific combination of characteristics. We have not observed any dependencies reported among different evaluated characteristics in the reviewed literature. Namely, all characteristics were evaluated separately, and the quality of a certain characteristic was never put in relation with the quality of another one.

2.4.4 Design of Evaluation

Extrinsic and *intrinsic* evaluation are two different types of human evaluation. Extrinsic evaluation assesses the ability of the system to perform an over-arching task with a real-world application. On the other hand, intrinsic evaluation assesses specific qualities or attributes and is evaluated independently of the over-arching task. Therefore, a system could perform well intrinsically without performing well extrinsically. Most papers (88%, 153/173) performed intrinsic evaluation, 4% (7/173) performed extrinsic evaluation, and 8% (13/173) involved aspects of both intrinsic and extrinsic evaluation. Intrinsic evaluation remains popular likely due to its simplicity, cost-efficiency, ease in tracking progress and benchmarking (Gehrmann et al., 2023), (Belz and Gatt, 2008). The lack of extrinsic evaluation may also be affected by the difficulty of designing an evaluation that effectively emulates its usage in the real-world setting.

Bias mitigation is important due to the potential compromise of human evaluation caused by order effects (Van Der Lee et al., 2019). Order effects include practice, carryover, and fatigue effects (Van Der Lee et al., 2019), all of which have the potential to affect human evaluation and lead to misleading and biased results. To mitigate this, Van Der Lee et al. (2019) suggested potential solutions including practice trials, increasing the time between tasks, shortening tasks, and proposed specific evaluation designs such as counterbalancing (systematically varying the order of presentation) and randomization. Further solutions include multiple evaluators assessing the same point (Son et al., 2022) to increase the reliability of their human evaluation and randomized counterbalancing, which is a combination of randomization and counterbalancing methods (Kurisinkel and Chen, 2019). However, the method of bias mitigation was only specified in 14% (24/173) of papers. This may be due to the high costs of evaluation designs, specifically counterbalancing. However, according to Van Der Lee et al. (2019), randomization or limiting the evaluation to one judge per system (if order effects are suspected) should be sufficient to mitigate order effects and avoid biased results.

2.4.5 Statistical Analysis

Inter-annotator agreement (IAA) scores should be reported to confirm consistency between evaluators and the reliability of the evaluation. Typically, a higher score indicates increased IAA. 34% of included papers (58/173) reported IAA using Kendall's τ, Fleiss' κ, Cohen's κ, Krippendorf's α and percentage agreement to name a few. However, a detailed analysis of the IAA scores and how they affected the overall evaluation is important. In some cases, IAA scores can prove to not be a useful measurement of agreement - as alluded to further in (Amidei et al., 2018b).

The importance of ensuring the reliability and validity of human evaluation is further highlighted by Liu et al. (2022) through the need for using statistical tests. Other methods of presenting data and analyzing results include displaying 1st and 2nd best performances in a table by highlighting the specific performance values (Gangal et al., 2022); or summary statistics such as standard deviations or mean scores (Qian and Levy, 2022). Only 16% of papers (28/173) used statistical tests as a form of analysis of their human evaluation such as student's t-test and Wilcoxon ranked test (Van Der Lee et al., 2019). This could be due to a lack of statistical power attributed to inadequate sample sizes, which could lead to misleading or different conclusions as they are more subject to the effects of chance (Otani et al., 2023).

3 Hierarchical Evaluation Framework

The review of existing literature identified 3 gaps:

- Majority of human evaluation was *intrinsic*.
- The characteristics of NLP systems were evaluated *independently*.
- Human evaluation focused on assessing the *outputs* of NLP systems, neglecting the evaluation of their *inputs*.

The analysis of existing literature revealed that the majority of papers (88%, 153/173) focused solely on an intrinsic evaluation of NLP systems. To avoid conducting an evaluation merely for the sake of it, we suggest that first a clear purpose for an NLP system is defined, and subsequently, an extrinsic evaluation is designed to gauge the systems' performance in fulfilling that specific purpose.

Additionally, the evaluation of various aspects of NLP systems' outputs (e.g., truthfulness) is usually conducted independently, without providing a composite score for the overall system performance. We suggest adopting a hierarchical approach, where the characteristics of the systems are interdependent, and the evaluation process continues only if the preceding characteristic(s) is deemed satisfactory. Conversely, if a characteristic is unsatisfactory, the evaluation can be discontinued, allowing evaluators to save time by not evaluating all characteristics for the low-quality outputs.

Lastly, to date, the existing literature has focused solely on the human evaluation of NLP systems' outputs, assuming that the inputs provided to these systems were of good quality. However, this assumption may not always hold true. We thus propose a two-phase approach for human evaluation, wherein testers initially assess NLP systems, followed by evaluators who evaluate both the inputs and outputs of the systems. By dividing the evaluation process into two phases, we enable evaluators to also assess the quality of the inputs used by testers during the testing phase of NLP systems. In essence, our hypothesis is that the quality of the outputs may not only be influenced by the system itself but also by the quality of the inputs.

In order to address those gaps, we propose a framework as shown in Figure 4. By defining a system's purpose as the first step, our framework supports extrinsic evaluation. The second step is to define interdependencies between the evaluated characteristics and consequently to design a hierarchical evaluation metric that supports calculating a composite score that encompasses the overall quality of an NLP system. Namely, the evaluation stops if any of the evaluated characteristics is deemed unsatisfactory and, in this case, the composite score is "bad" as the system did not pass the evaluation. Otherwise, if the evaluation goes to the end, then the composite score is "good". We hypothesize that our framework facilitates a shorter evaluation time for evaluators by allowing early termination of evaluation in cases where any evaluated characteristic does not meet satisfactory quality. The third step is to do testing of the system according to the defined purpose. Testers are independent of evaluators who evaluate the system's inputs and outputs using the designed hierarchical evaluation metric in the fourth step. This allows for independent evaluation of the system's inputs as well. Consequently, our framework enables an examination of whether the quality of a system's outputs is influenced by the quality of its inputs.

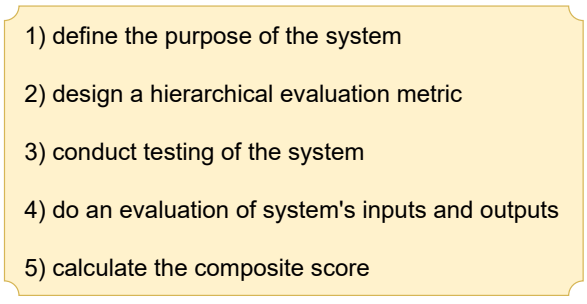

Figure 4: Steps explaining how to create a hierarchical evaluation framework for an NLP system.

4 Case study: Hierarchical Evaluation for an MRC System

We evaluated a Machine Reading Comprehension (MRC) system using the framework outlined in the previous section. In an MRC system, answers come in the form of short text spans which are directly extracted from the text corpus (i.e., relevant text database). Questions asked, on the other hand, need to be relevant to the topic that the text corpus covers, factoid, answerable and mistake-free (i.e., no spelling or grammar mistakes).

4.1 The purpose of the MRC System

The purpose of the developed MRC system was to support health coaches during their sessions with clients, coaching them on the importance of good quality sleep. Namely, the developed system is part of the human-AI symbiosis model shown in Figure 5 (Bojic et al., 2023b). The system is a pre-trained BERT model that was fine-tuned on a human-annotated domain-specific dataset.

The entire health coaching process takes place online through text messaging. To address factoid questions raised by clients, the health coach may utilize the MRC system for additional support during coaching sessions (Bojic et al., 2022, 2023a). Health coaches were given the liberty to use, modify, or disregard the answers provided by the MRC system. This integration enhances the human coaching experience by incorporating evidence-based knowledge given by the MRC system. As a result, the health coaches' response time improves, and the information they offer is grounded in reliable evidence.

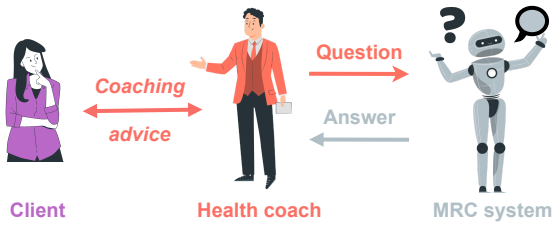

Figure 5: Human-AI health coaching model.

4.2 Hierarchical Evaluation Metrics

We developed two evaluation metrics: one for the inputs (i.e., questions) of the MRC system and the other for the outputs (i.e., answers), in order to be able to detect whether the quality of the MRC system output is affected by the quality of its input.

4.2.1 Evaluation of Inputs

Figure 6 shows a set of evaluation criteria for evaluating the MRC questions. The question is *relevant* if it is on the topic covered in the corresponding text corpus. *Factoid* questions are questions that start with one of the following words: "who", "what", "where", "when", "why" or "how". They ask about facts that can be expressed as short texts (Parsing, 2009). The question is *answerable* if there exists an answer to it. The evaluators are asked if the posed question contains any *spelling* or *grammar* errors. The *difficulty* of the posed question can be chosen from three levels – *easy*, *medium*, or *hard* (please refer to Table 1).

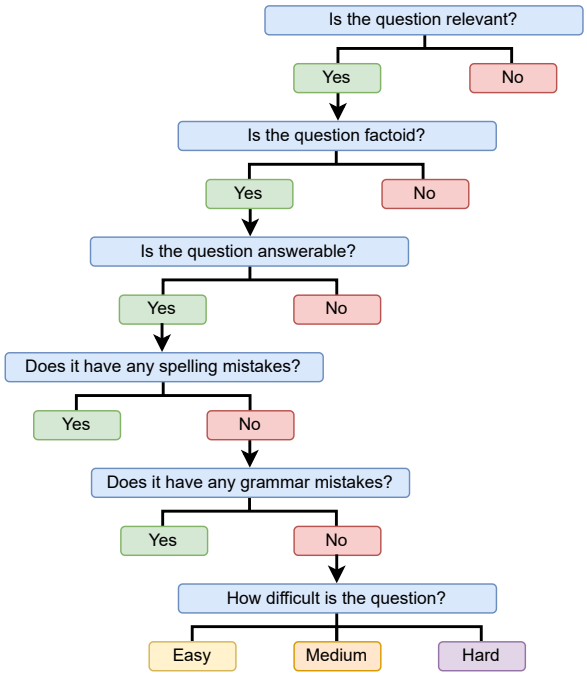

Figure 6: Hierarchical evaluation of the questions.

Table 1: Three different levels of difficulty of the posed questions.

Easy	The correct answer is obvious after reading the passage only one time.
Medium	To find the correct answer, one needs to carefully read and understand both the question and the paragraph.
Hard	To find the correct answer, one needs to read the paragraph many times, sometimes even use logical reasoning to find the correct answer.

4.2.2 Evaluation of Outputs

The evaluators were asked to evaluate the retrieved *short answer* and if necessary its *explanation*. Namely, the output of the whole MRC system is a text span (i.e., short answer). However, an MRC system can be seen as a pipeline of two NLP models - *document retrieval* and *document reader*, where the output of the former model is the *relevant passage(s)* and the output of the latter model (i.e., the whole system) is a *text span*. Our metric first evaluates the characteristics of the output of the whole system (i.e., text span). If the output of the whole system was not satisfying, then we evaluate its explanation (i.e., relevant passage) that was provided by the document retrieval component.

The retrieved short answer is *clear* if its meaning is easy to understand. The retrieved short answer/explanation is *relevant* if it answers the posed question. *Clinical accuracy* of the retrieved short answer/explanation denotes the degree to which it is clinically accurate – (i) clinically accurate, (ii) partially clinically accurate, and (iii) clinically inaccurate (see Table 2). Finally, the health coaches judged the usefulness of the retrieved short answer/explanation (see Figure 7).

Table 2: Three different levels of clinical accuracy.

Clinically accurate	The retrieved short answer/ explanation is clinically accurate and is based on evidence-based information.
Partially clinically accurate	The retrieved short answer/ explanation is partially clinically accurate and somewhat lacks evidence-based information.
Clinically inaccurate	The retrieved short answer/ explanation is not clinically accurate and is not based on evidence-based information.

4.3 Testing of the MRC System

Testing of the developed MRC system was conducted during a pilot Randomized Controlled Trial (RCT). In this RCT, 30 participants in the intervention group (i.e., clients) interacted with 10 health coaches who utilized the MRC system to answer factoid questions. Clients were recruited from a general student population if they (1) were older than 21 years, (2) were available for weekly interaction with a health coach for four weeks, (3) were not currently undergoing any treatment for a sleep disorder or mental disorder and were not under the care of a psychologist or psychiatrist, and (iv) had PHQ-9 score less than 10.

Health coaches were recruited from the cohorts of graduated students from the health coaching course if they (1) were older than 21 years, (2) were available for weekly interaction with three clients for four weeks, and (iii) successfully completed and passed the health coaching course. During the study period of four weeks, clients had weekly 30-minute sessions with their respective health coaches. All questions asked by health coaches and their corresponding answers were saved during the testing phase and were subsequently used in the evaluation phase. By dividing human evaluation into two parts, we were able also to judge whether questions were posed in the way we asked our health coaches to ask them, i.e., if they can be answered by the developed MRC system.

4.4 Evaluation of the MRC System

Following a 4-week pilot RCT, the developed MRC system underwent evaluation by 10 health coaches. A total of 387 unique question-answer pairs were evaluated by the health coaches during this period. The heat map depicted in Figure 8 illustrates the number of inputs and outputs evaluated by each health coach, while Figure 9 showcases the average evaluation time required for each input/output assessed by the health coaches.

Almost all questions (99%, 383/387) were evaluated as *relevant*. One example of a question that was marked as not relevant was: "*Food nutrition tips*". The next 87% of questions (335/383) were judged as factoid. Some examples of not factoid questions are as follows: "*About REM sleep, is it the phase that I'm dreaming?*", "*Can you exercise before sleeping?*", "*I often run around campus for 3-5km at night 1-2h before sleeping. Is it good or bad for sleep?*". 2% of the remaining questions (8/335) were marked as not answerable: *"How long should I be awake during sleep?"*, *"How bad would you say is my sleep health like compared to the average?"*, while additional 2% (6/327) had spelling errors (e.g., *"How long before bedtime shld i stop screentime?"*). Finally, the last 23% (74/321) had grammar errors: *"How do ensure naps have good quality?"*, *"Why wake up during night?"*. The results of the complete external human evaluation for questions are shown in Figure 10.

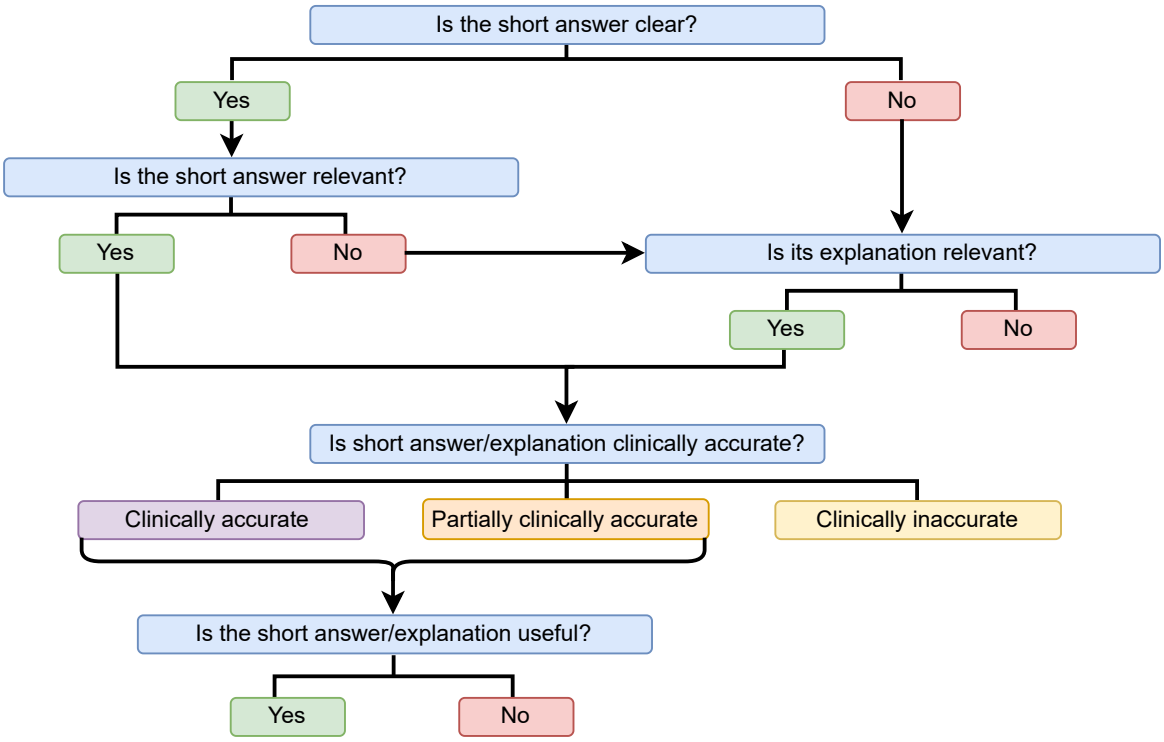

Figure 7: Hierarchical evaluation of the answers.

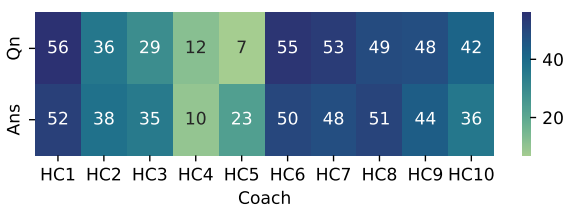

Figure 8: The total number of questions and answers evaluated by each health coach.

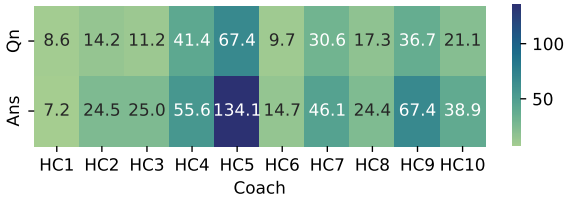

Figure 9: Average time in seconds per health coach needed to evaluate questions and answers.

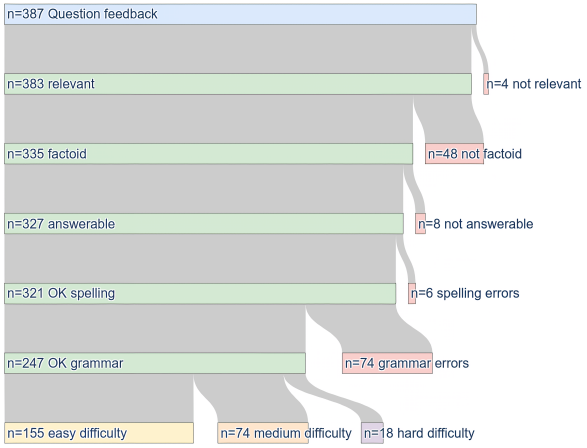

Figure 10: Extrinsic evaluation of questions.

More than 40% (157/387) of short answers were evaluated as not *clear*, out of which in 57% of cases (89/157), their explanations were marked as relevant. For example, "***Question**: When does melatonin peak? **Answer**: release of melatonin, the hormone that induces feelings of tiredness and relaxation. **Explanation**: When the sun goes down, your eyes will perceive darkness and signal the scn accordingly. This triggers the release of melatonin, the hormone that induces feelings of tiredness and relaxation. This also causes your core temperature to dip.*". 63% of clear answers (146/230) were also evaluated as relevant of which 99% (144/146) was indicated as being (partly) clinically accurate. Furthermore, 97% (113/116) of the short answers that were not clear, but their explanations were relevant, were (partly) clinically accurate. The results of the complete external human evaluation for answers are shown in Figure 11.

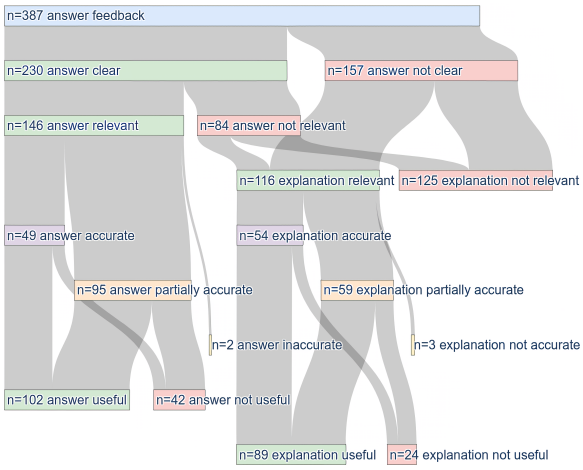

Figure 11: Extrinsic evaluation of answers.

4.5 Composite scores of the MRC System

The results of our evaluation showed that 63.8% (247/387) of unique questions were evaluated as relevant, factoid, answerable, spelling and grammar mistakes-free (i.e., *good* questions). Out of those, 63% (155/247) were judged as easy, 30% (74/247) as medium and 7% (18/247) as hard questions. Furthermore, 49.4% (191/387) of unique answers were evaluated as clear, relevant, clinically accurate and useful (i.e., *good* answers). In order to check if there are any associations between the quality of outputs and inputs, we performed a χ^2 test. The result showed significant associations between the two (χ^2 = 4.56, p=0.03). The distribution of the performance matrix is shown in Table 3.

Table 3: 2x2 matrix for the performed χ^2 test.

		Questions	
		good	bad
Answers	good	132	59
	bad	115	81

5 Discussion and Conclusions

In this study, we conducted a scoping review to identify gaps in the literature regarding human evaluation in NLP. The findings revealed three significant gaps that need to be addressed: the lack of evaluation metrics for NLP system inputs, limited consideration for interdependencies among different characteristics of NLP systems, and a scarcity of metrics for extrinsic evaluation.

To bridge these gaps and enhance human evaluation in NLP, we proposed a hierarchical evaluation framework. Our framework offers a standardized approach that considers both the inputs and outputs of NLP systems, allowing for a more comprehensive assessment. Moreover, our hierarchical approach considers the interdependencies among different characteristics of NLP systems. Rather than evaluating characteristics independently, our framework emphasizes their interconnectedness and the impact they may have on each other. This approach enables a more holistic evaluation that captures the overall performance of NLP systems.

To validate the effectiveness of our proposed framework, we conducted a pilot RCT evaluating an MRC system. The evaluation phase of our study involved 10 health coaches who evaluated a total of 387 question-answer pairs generated during the RCT. The evaluation metrics developed for inputs focused on aspects such as relevance, factoid nature, answerability, spelling, grammar errors, and difficulty levels of the questions. For outputs, the evaluation criteria included clarity, relevance, clinical accuracy, and usefulness of the retrieved short answers and explanations.

The results of the evaluation provided valuable insights into the strengths and weaknesses of the MRC system and demonstrated the practical application of our hierarchical evaluation framework. The findings supported the notion that evaluating both inputs and outputs is crucial for obtaining a comprehensive understanding of the performance and effectiveness of NLP systems. Future research should focus on validating the scalability and time-saving benefits of our proposed framework.

Limitations

We recognize the potential limitations that may arise with a small-scale scoping review that is limited to a few venues. As our sample size is small, our results and proposed solutions may lack generalizability and applicability. To mitigate the potentially negative effects, we carefully chose the most appropriate venues - as further explained in 2.1 - and limited the search to the most recent papers as the field of computer science is rapidly and constantly evolving. Solely reviewing papers in the English language could also potentially limit the scope of our research. We also tried to delve into a broad range of aspects of human evaluation whilst keeping our objectives focused. However, we recognize the inevitability of potential factors that may exist outside of our considerations - which may also affect results and conclusions.

Ethics Statement

We aim to conduct our study with the highest ethical standards and maintain continuous referral to the ACL code of ethics throughout our research. We obtained articles via Google Scholar and have anonymized most of the papers and authors - excluding a few that were cited in our main text. This paper should be used to provide insight into the current practices of human evaluation and a potential solution to streamline the process. It is not used to penalize any research or draw any negative attention to certain papers.

We also recognize that some potential biases and errors may arise amongst human reviewers which may lead to potentially inaccurate data extraction. This may have a potential knock-on effect on derived conclusions. These issues are considered and mitigated through multiple reviewers performing the same task, frequent discussions, and good communication.

Acknowledgements

The authors would like to acknowledge the Accelerating Creativity and Excellence (ACE) Award (NTU-ACE2020-05) and center funding from Nanyang Technological University, Singapore. Josip Car's post at Imperial College London is supported by the NIHR NW London Applied Research Collaboration. Finally, the authors would also like to acknowledge Jintana Liu and Ashwini Lawate who were included in the pilot RCT running and supported the data collection process.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018a. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018b. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200.

Iva Bojic, Josef Halim, Verena Suharman, Sreeja Tar, Qi Chwen Ong, Duy Phung, Mathieu Ravaut, Shafiq Joty, and Josip Car. 2023a. A data-centric framework for improving domain-specific machine reading comprehension datasets. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 19–32, Dubrovnik, Croatia. Association for Computational Linguistics.

Iva Bojic, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023b. Building extractive question answering system to support human-ai health coaching model for sleep domain. *arXiv preprint arXiv:2305.19707*.

Iva Bojic, Qi Chwen Ong, Megh Thakkar, Esha Kamran, Irving Yu Le Shua, Jaime Rei Ern Pang, Jessica Chen, Vaaruni Nayak, Shafiq Joty, and Josip Car. 2022. Sleepqa: A health coaching dataset on sleep for extractive question answering. In *Machine Learning for Health*, pages 199–217. PMLR.

Sheila Castilho. 2021. Towards document-level human mt evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Association for Computational Linguistics (ACL).

Varun Gangal, Steven Y Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10645–10653.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definition. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics (ACL).

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96.

Litton J Kurisinkel and Nancy Chen. 2019. Set to ordered text: Generating discharge instructions from medical billing codes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6165–6175.

Hongru Liang and Huaqing Li. 2021. Towards standard criteria for human evaluation of chatbots: A survey. *arXiv preprint arXiv:2105.11197*.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981*.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286.

Patrick Paroubek, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of evaluation in natural language processing. In *Traitement Automatique des Langues, Volume 48, Numéro 1 : Principes de l'évaluation en Traitement Automatique des Langues [Principles of Evaluation in Natural Language Processing]*, pages 7–31, France. ATALA (Association pour le Traitement Automatique des Langues).

Constituency Parsing. 2009. Speech and language processing. *Power Point Slides*.

Micah DJ Peters, Christina M Godfrey, Patricia McInerney, Cassia Baldini Soares, Hanan Khalil, and Deborah Parker. 2015. *The Joanna Briggs Institute reviewers' manual 2015: methodology for JBI scoping reviews*, chapter 11. The Joanna Briggs Institute.

Peng Qian and Roger Levy. 2022. Flexible generation from fragmentary linguistic input. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8176–8196.

Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary korean and english. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1260–1272.

Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. 2021. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 46–55.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Yan Zeng and Jian-Yun Nie. 2021. An investigation of suitability of pre-trained language models for dialogue generation–avoiding discrepancies. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4481–4494.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation. *arXiv preprint arXiv:2004.10450*.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3438–3448.

Appendix

Appendix 1: Data Extraction Form

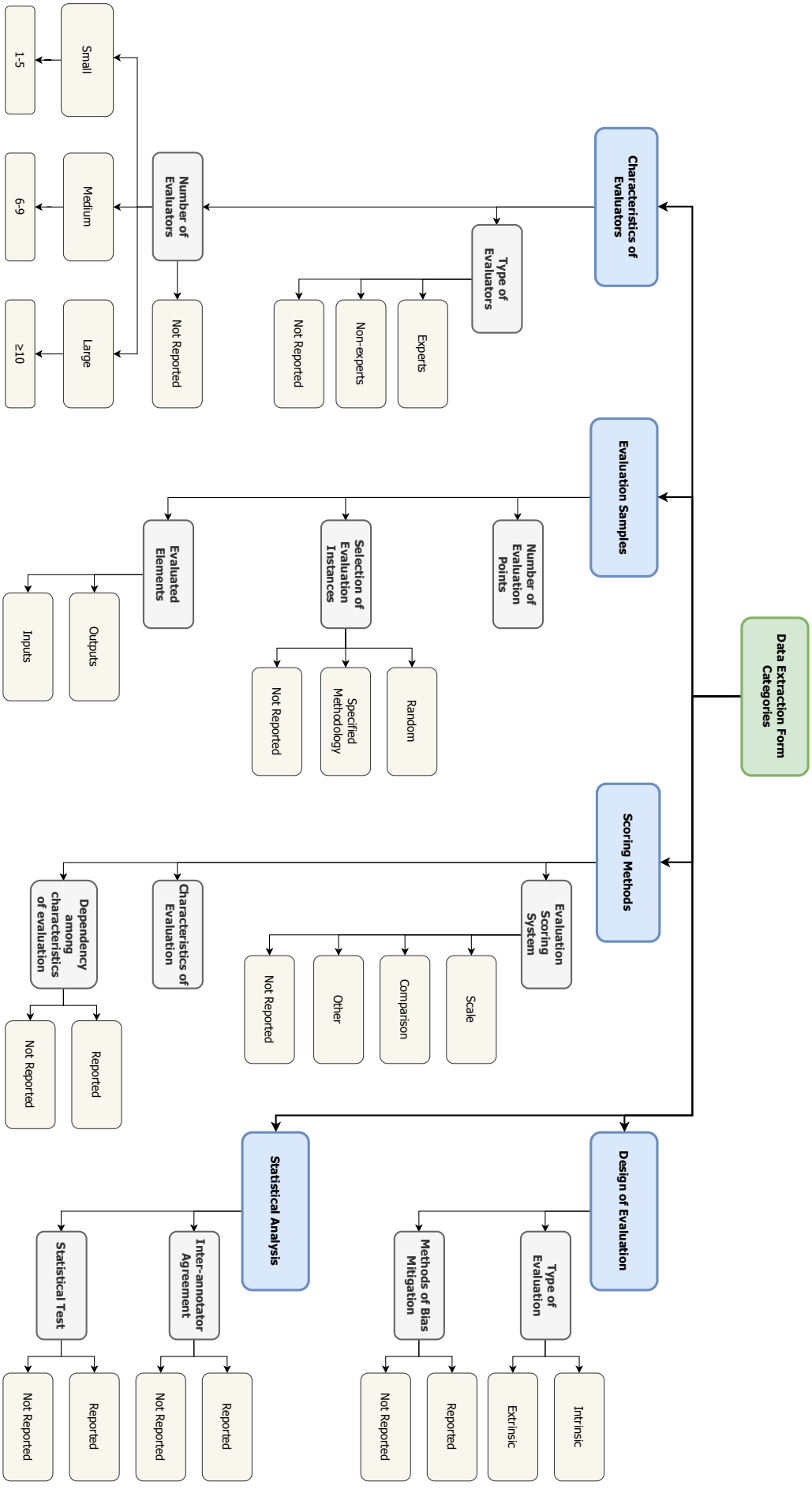

Figure 12: Data extraction form categories.