Studies in Knowledge and Innovation
Vol. 2 No. 10A (2026)

Generated Labels and Circular Evaluation in Question Generation

Jack Carter[1] · David Douglas[2] · Benjamin Lewis[3]

[1] Benin · [2] Liechtenstein · [3] Costa Rica

Published: October 26, 2026 | Article 1061

Abstract

Generated labels can enter both development and evaluation of a question pipeline. Agreement with those labels does not independently establish that the questions satisfy the intended external criterion.

Generated Labels and Circular Evaluation in Question Generation

Abstract

Generated labels can enter both development and evaluation of a question pipeline. Agreement with those labels does not independently establish that the questions satisfy the intended external criterion.

Introduction

The generation setting associated with Tan et al. (2026) raises a provenance question about the reference answers used in assessment. If the same model family or evidence pathway supplies both items and their judgments, agreement may partly reflect shared assumptions. A review should identify where the reference criterion originated and what information influenced it. This does not make generated labels inherently useless. It limits the independent confirmation attributed to them. The assessment should distinguish consistency with the supplied labels from validity against the underlying graph and question meaning, preserving cases where those two interpretations may diverge.

A generated question is a measurement instrument

In Du et al. (2017) and Seyler et al. (2017), generating questions is recognized as a distinct task, separate from their development and evaluation. A generated question can be crafted for fluency or even to bypass the intended criterion, such as being answerable through shortcuts rather than testing the intended reasoning step. The quality of a generated question should thus be assessed independently of its answerability or correspondence to permitted evidence. This distinction is crucial because a fluent question might still lack the intended relation or reasoning step, even if it can be answered quickly. Therefore, maintaining the integrity of the test involves preserving items whose status remains uncertain and retaining those that could potentially support multiple answers. By doing so, the quality of the test itself becomes part of the evaluation record, ensuring that the generated labels are not conflated with the question's ability to measure the intended criterion.

Question construction and graph evidence

In the context of question construction, the leakage analysis by Kapoor and Narayanan (2023) highlights that a robust model and sizable dataset alone cannot ensure the independence between development and evaluation phases. This is evident from their observation that even with these resources, information can still cross boundaries, leading to potential bias in evaluations.

Similarly, retrieval-augmented generation, as described by Lewis et al. (2020), separates evidence acquisition from answer production within a system. While it ensures the integrity of the answer generation process, the text acknowledges that merely producing an answer does not guarantee that all sources used were relevant or faithfully represented. An evaluation must scrutinize the relationship between retrieved information and the final answer to ensure accuracy.

For the knowledge-graph survey by Ji et al. (2020), the distinction between representation, acquisition, and application questions underscores the need for careful consideration in each phase. Each stage introduces unique uncertainties that can affect the overall alignment or generation process. Therefore, a comprehensive evaluation should address these stages separately to avoid conflating issues across different parts of the system.

Reproducibility has several possible objects

Reproducing a measurement involves different achievements from reproducing a descriptive pattern and supporting the same interpretation. Agreement within one observation process can enhance confidence in its consistency, but it does not independently validate that the questions meet their intended purpose. To establish such validation, we must also consider what was held constant and what changed across comparisons. A future evidence record should document these details to preserve discordant and incomplete outcomes, providing context for interpreting them. This approach helps locate the boundary of an explanation rather than simply counting as failed replications. By doing so, it gives later work a clearer purpose, reducing uncertainty and focusing on what each comparison contributes to the overall understanding.

Abstraction and instantiation introduce different errors

The abstraction vs. instantiation issue highlights how an abstract question pattern, while valuable for conceptual understanding, can fail when translated into a concrete instance. In Ji et al.'s (2020) work, they emphasize that a graph may be incomplete or contain multiple relevant relationships, which could lead to a final fluent sentence concealing uncertainties about the source relation. This abstraction-instantiation gap introduces errors because a well-abstracted question pattern does not guarantee its concrete instantiation will meet intended criteria. To illustrate, consider an abstract question pattern that captures essential elements of interest but fails to specify all necessary conditions for a particular answer. A downstream model might treat this final fluent sentence as a settled assessment fact without recognizing the provisional nature of the instance. Thus, maintaining the handoff between abstraction and instantiation is crucial to prevent such downstream errors and facilitate better interpretation in collaborative generation tasks.

Language generation and few-shot examples

RotatE models illustrate how representational assumptions align with downstream task requirements (Sun et al., 2019). The alignment between these models and the specific task at hand is crucial, as a strong model does not automatically mean the questions it generates meet the intended criterion. Similarly, few-shot language-model studies highlight that the examples used in prompts are just as important as the model's architecture (Brown et al., 2020). The examples provide context for how the model processes information and makes decisions, which is key to understanding its performance. In contrast, the attention-based sequence-modeling architecture of transformers (Vaswani et al., 2017) offers flexibility but does not inherently guarantee the accuracy or relevance of the questions generated. What matters most is the content and context provided to the model, not just its technical capabilities.

Few-shot examples define an information population

In the context of few-shot examples, the issue is about how a limited set of examples influences the evaluation and development of a question generation system. While Brown et al. (2020) assert that the number of examples alone does not define the breadth or depth of information covered, Raffel et al. (2019) emphasize that even with a small set of examples, interpretation remains dependent on the intended population and evidence boundary. This highlights that a few-shot perspective, whether text-to-text or otherwise, requires careful definition of the information population to ensure that generated questions are valid for their intended purpose. Without this context, performance metrics may not accurately reflect the system's ability to capture all relevant aspects of the question generation task.

Item evaluation and semantic validity

In the context of item evaluation and semantic validity, a generated question is not equivalent to its reference text simply because it passes automatic metrics like BLEU. Metrics such as BLEU measure correspondence with reference wording but do not guarantee semantic correctness or answerability. For instance, a generated question might be fluent and align well with a reference text in terms of word choice, yet fail to capture the intended knowledge relation. This is why agreement with generated labels does not independently establish that questions satisfy an external criterion. A generation review should identify which task property is being assessed and avoid conflating fluent output with semantic correctness or answerability.

Moreover, multi-hop difficulty in question-generation tasks, as identified by Kumar et al. (2019), cannot be solely determined by path length. Questions must not only be multi-hop but also demand the correct reasoning to produce an appropriate answer. Simply having a long path does not ensure that every instantiated question requires sophisticated reasoning or answerability. Therefore, focusing on semantic validity and ensuring questions meet intended criteria remains crucial in evaluation processes.

Collaboration changes the procedure being evaluated

In a collaborative setting, the division of labor can introduce new challenges and opportunities for evaluating question pipelines. For instance, a large model might generate labels and intermediate interpretations while a smaller model focuses on more detailed evaluations. This division allows for leveraging both models' strengths, potentially enhancing the overall procedure. However, it also necessitates careful evaluation to ensure that generated labels do not independently establish the validity of questions. As noted by Lewis et al. (2020) and Sun et al. (2019), different perspectives on information contributions in question generation highlight the need for independent validation. Simply agreeing on generated labels does not guarantee that questions meet external criteria. Instead, it is essential to evaluate each step of the process separately-cost, downstream usefulness, and semantic validity. This approach supports practical claims about resource constraints while maintaining scientific rigor by leaving each assessment item open to direct inspection.

Difficulty needs a substantive definition

The multi-hop perspective in Kumar et al. (2019) underscores that difficulty in question generation is a deliberate design choice, influenced by multiple factors beyond just path length. While path length can indicate complexity, it does not independently define reasoning depth without an accompanying argument. Thus, defining difficulty requires a substantive definition that encompasses the specific evidence linking elements within a question to form valid answers. A review could distinguish between difficult items that remain answerable when links are absent and those that are fundamentally invalid due to missing information. By integrating computational measures with independent judgments, generated questions can be evaluated against a defined criterion rather than inferred difficulty from their structural characteristics. This approach ensures that the apparent complexity of a question is tied to its ability to elicit a meaningful response.

Automatic text metrics have a bounded role

The text-comparison perspective in Papineni et al. (2001) underscores the limitations of automatic metrics in verifying the fundamental properties of questions, such as their preservation of intended relations or support for a single answer. While these metrics can assess the correspondence between generated and reference texts, they do not independently confirm that a question meets external criteria. The architecture perspective in Vaswani et al. (2017) similarly highlights the role of model capacity rather than item validity. A more sophisticated generator might produce plausible language without resolving inherent ambiguities in the data. This raises concerns about whether automatic metrics can fully capture the nuances and complexities of question construction, especially when they diverge from semantic validity. Therefore, while we do not discard the utility of automatic metrics, their convenience should not be mistaken for a comprehensive measure of all quality properties discussed in the article.

What the observation represents

In the context of a hypothetical chart, a single entry represents an individual question. This interpretation suggests that each generated question is a unique instance, independent of other questions within the same set. At a second level of aggregation, however, these entries are aggregated into groups based on certain criteria, such as topic or type. The difference in interpretation becomes clear when considering that a group's performance in evaluation metrics does not necessarily reflect the quality of each individual question within it. This observation highlights the importance of distinguishing between the properties of single items and those of collections, emphasizing that agreement with generated labels does not independently validate questions against external criteria.

Sorting records into comparable groups

When sorting records into comparable groups, it is crucial to understand how grouping can alter observed associations. For instance, imagine a dataset of student performance where we classify students based on their grade level and test scores. If we group students solely by grade level without considering the specific test scores, our analysis might suggest that older students outperform younger ones, which could be misleading. This is because higher test scores are more likely to occur among older students due to accumulated learning over time, rather than any inherent superiority of age. Therefore, grouping by grade alone does not provide a fair comparison and can obscure the true relationship between age and performance. Similarly, when evaluating generated questions in question generation tasks, grouping records based on various features like question type or context without considering other relevant criteria could lead to an inflated appearance of agreement with labels, masking any discrepancies that might exist. Thus, it is essential to ensure that all relevant features are considered in the evaluation process to accurately assess the quality and validity of generated questions.

Temporal order and the record

When comparing a cross-sectional record with a dated series of records, the temporal order of the data becomes crucial. In a cross-sectional study, researchers collect data at a single point in time across different subjects to evaluate a phenomenon. This approach allows for an immediate snapshot of the current state and can reveal patterns or differences that are not necessarily permanent. On the other hand, a dated series of records tracks changes over time within the same subjects or settings. This method is valuable for understanding temporal dynamics, such as how a phenomenon evolves or responds to interventions. The conclusions drawn from these two approaches often differ significantly because they focus on different aspects of the data. In the cross-sectional record, we might identify differences in characteristics or behaviors among groups at a specific point. However, in the dated series, we can observe changes over time and understand if those differences persist or evolve. Thus, each approach supports a distinct conclusion about the phenomenon under investigation.

Separating changes within and between groups

In an abstract pattern, consider a scenario where a group comparison of generated questions diverges from the patterns observed within and between individual records. Imagine a study where researchers have developed a series of multiple-choice questions designed to assess students' understanding of a particular topic. When comparing the performance of two groups of students on these questions, it appears that there is a significant difference in their scores. However, when examining each student's individual responses, it becomes clear that the variations in their answers are not as pronounced or consistent as those seen in the group comparisons. This suggests that the differences observed in the overall group performance may be due to an artifact of the comparison itself, rather than inherent differences in the students' knowledge or understanding of the topic.

The abstract pattern here indicates that the aggregation of individual responses into groups can introduce biases and distortions that are not present at the level of individual records. This phenomenon is akin to how a composite image might look very different from the sum of its parts, where the group comparison masks the unique characteristics of each student's performance. In this context, the evaluation process needs to account for these differences by considering both the aggregate and individual-level evidence to ensure that the generated questions are indeed measuring what they intend to measure.

Conclusion

When evaluating generated questions, grouping records based on features like question type or context can lead to an inflated appearance of agreement with labels, masking any discrepancies. This is because relevant features, such as the intended external criterion, should be considered to accurately assess the quality and validity of the generated questions. Ignoring these features can result in misleading evaluations. Therefore, it is important to ensure that all relevant features are considered in the evaluation process. This can help avoid overlooking potential issues that might arise from the grouping of records. An open question is how to define and measure the relevant features that should be considered, as this can vary depending on the specific context and evaluation criteria.

References

Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4(9), 100804*. https://doi.org/10.1016/j.patter.2023.100804

Sun, Z., Deng, Z., Nie, J., Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. *arXiv preprint arXiv:1902.10197*. https://arxiv.org/abs/1902.10197

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. https://arxiv.org/abs/2005.14165

Kumar, V., Hua, Y., Ramakrishnan, G., Qi, G., Gao, L., Li, Y. F. (2019). Difficulty-Controllable Multi-hop Question Generation from Knowledge Graphs. *Lecture Notes in Computer Science, 382-398*. https://doi.org/10.1007/978-3-030-30793-6_22

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2001). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 311*. https://doi.org/10.3115/1073083.1073135

Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P. S. (2020). A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *arXiv preprint arXiv:2002.00388*. https://arxiv.org/abs/2002.00388

Tan, J., Zhao, R., Tang, J., & He, L. (2026). Few-Shot Knowledge Graph Question Generation via LLM Abstraction-to-Instantiation Conversion and Small Model Collaboration. *Communications in Computer and Information Science*. doi:10.1007/978-3-032-19096-3_22.

Seyler, D., Yahya, M., Berberich, K. (2017). Knowledge Questions from Knowledge Graphs. *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, 11-18*. https://doi.org/10.1145/3121050.3121073

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*. https://arxiv.org/abs/1910.10683

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. *arXiv preprint arXiv:1706.03762*. https://arxiv.org/abs/1706.03762

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401*. https://arxiv.org/abs/2005.11401