University of Illinois Chicago (UIC)
Recruitment Advertisement for Participation in Research

STUDYID: STUDY2025-0188

Plain Text Summary Human Evaluation Guidelines

Your goal of this evaluation is to assess the quality of model-generated summaries. You will rate the summaries based on their comprehensiveness, layness, usefulness, and factuality.

Materials Provided for Evaluation:

- Abstract of the Article
- Model-generated plain text for the abstract

Evaluation Criteria:

- We will use a 1-5 Likert scale where 1 indicates poor quality and 5 indicates excellent quality.
- Below are the individual sores meaning for each of the five facets

Comprehensiveness:

Aim to assess how well the model output contains information necessary for a non-expert to understand the high-level topic and significance of the research. Explanation for each rating is as follows:

1. The summary is incomplete, an evaluator cannot understand the topic or the significance of the research.
2. The summary is partially complete; an evaluator gains a vague idea of the topic but cannot grasp the significance due to missing key details.
3. The summary allows an evaluator to understand the topic but lacks important details that convey the research's significance.
4. The summary enables an evaluator to understand both the topic and significance, missing only minor details that could enhance understanding.
5. The summary thoroughly covers all necessary information, allowing an evaluator to fully understand the topic and the significance of the research.

Layness

We measure it based on medical jargon, sentence structure, and explanations for the terms in the plain text. Explanation for each rating is as follows:

1. There is not much difference between the plain text summary and the original abstract.
2. The plain text summary omits a few sentences that include jargon or omits a few words in sentences. It becomes easier to read but does not truly simplify the content.
3. The summary is a mix of jargon and simple terms, as well as simple and complex sentences, along with some definitions. Laypersons may understand the main points but could find specific terms or sentences confusing.
4. The summary is overall easy to understand, with the occasional presence of a complex sentence or medical terms that are not explained to the reader.
5. The summary removes jargon or uses simple synonyms for them. If it cannot do either, it adds context for the evaluator to grasp the complex term. It uses simple, straightforward sentences or makes use of examples, making it easy for anyone to understand.

Factuality

We measure factuality as to what extent the plain text is factually consistent with the abstract. *Specifically, by looking at the information stated in the abstract. (Intrinsic Hallucination)* Explanation for each rating is as follows:

1. The study alters the findings or methodology, misrepresenting the study.
2. The study alters part of the study that can lead to misinterpretation of sections such as method or results, but not the entire study.
3. The summary contains accurate information about the study but with frequent minor inconsistencies such as typos, incorrect figures, or omitting key details in findings.
4. The study contains accurate information about the study but with one or two minor exceptions.
5. The summary is fully factual and aligns completely with the study.

Usefulness

To evaluate if the summary is appropriately detailed for experts or simplified for the general audience, ensuring it matches the knowledge level of the audience (personas) and provides practical value. Explanation for each rating is as follows:

1. The summary fails to align with my knowledge and provides no value or useful information.

2. The summary partially aligns with my knowledge but provides limited value or information.

3. The summary adequately aligns with my knowledge, providing a fair amount of valuable information that somewhat meets my needs.
4. The summary aligns well with my knowledge, offering high-value and substantial information that meets most of my informational needs.
5. The summary perfectly aligns with my knowledge, providing maximum value and comprehensive information that fully satisfies my informational requirements.

Extrinsic Hallucinations

We also look for the presence of extrinsic hallucinations, specifically looking at the text added additionally by the model to enhance evaluator understanding. Unlike the Likert rating, we treat this measure as binary, where finding any one instance sets the value for the hallucination as true. We further classify it into three categories.

1. Contextual Inconsistency: When the model adds definitions, synonyms, or background context that are factually correct but contextually incorrect.
2. Factual Inconsistency: When the model adds definitions, synonyms, or background context that are factually incorrect but contextually correct.
3. Logical Inconsistency: When the model adds definitions, synonyms, or background context that are logically incorrect with respect to the context and the external knowledge.

An evaluator would also add their evaluation for extrinsic hallucination:

1. Contextual: True/False
2. Factual: True/False
3. Logical: True/False
4. Extrinsic Hallucinations: True/False (Based on the first three answers for inconsistencies, we can assign the value)

Procedure:

1. Each evaluator will independently review the abstract, the reference lay summary, and the model-generated summaries.
2. Evaluators will rate each summary based on the specified facets using the provided Likert scale.

Examples

Abstract 1:

Background: This study aimed to compare the arthroscopic internal drainage of popliteal cysts alone or in combination with cyst wall resection in terms of clinical outcomes. Methods: Forty-two consecutive patients with symptomatic popliteal cysts received arthroscopic treatment. Specifically, 20 of them received arthroscopic internal drainage (AI group) alone and 22 received arthroscopic internal drainage combined with cyst wall resection (AICR group) through double posteromedial portals. Magnetic resonance imaging (MRI) was performed to identify recurrence of popliteal cysts. The Lysholm score and Rauschning-Lindgren grade were used to assess the clinical outcomes. The median of the follow-up period was 24 months (12-48 months). Results: The two groups (AI group and AICR group) were similar in age, gender, cyst diameter, associated joint disorder, preoperative Lysholm score, preoperative Rauschning-Lindgren grade and follow-up period ($P > 0.05$). Relative to the AI group, the AICR group had a significantly prolonged operation time ($P < 0.05$) and a higher incidence of complications ($P < 0.05$). In both groups, the Rauschning-Lindgren grade at the last follow-up significantly differed from the preoperative grade ($P < 0.05$) and the Lysholm knee score remarkably increased compared to the preoperative score ($P < 0.05$); however, there were no differences between the two groups at the last follow-up ($P > 0.05$). According to the MRI results, the cyst disappeared in 11 (55%), shrank in size in 6 (30%) and existed in 3 (15%) patients in the AI group, and was absent in 18 (81.8%) and shrank in size in 4 (18.2%) patients in the AICR group, suggesting a significant difference between the two ($P < 0.05$). Conclusion: Additional resection of cyst wall can result in a lower recurrence rate of cysts but extend the operation time and increase the incidence of perioperative complications compared with arthroscopic internal drainage of popliteal cysts alone.

Model X:

This study aimed to compare the arthroscopic internal drainage of popliteal cysts alone or in combination with cyst wall resection in terms of clinical outcomes. Forty-two consecutive patients with symptomatic cysts received surgery to remove the cyst. Specifically, 20 of them received arthrosurgical internal drainage (AI group) alone and 22 received arroscopy internal drainage combined with cysts resection (AICR group) through double posteromedial portals. Magnetic resonance imaging (MRI) was performed to identify the recurrence of the cysts. The Lysholm score and Rauschning-Lindgren grade were used to assess the clinical results. The median of the follow-up period was 24 months (12-48 months). The two groups were similar in age, gender, cyst size, associated joint disorder, preoperative Lysholm scores, pre-operative Rausching-Lindberg grade and follow-ups period. Relative to the AI group, the AICR (AI) group had a significantly prolonged operation time and a higher number of complications. In both groups, the Lysholm scoring at the last followup significantly differed from the preoperative score. However, there were no differences between the two groups. In conclusion, additional resection of cyst walls can result in a lower recurrence rate of cysts but extend the operation time, increase the number of perioperative complications compared with the same procedure alone.

- Comprehensiveness: 3
  - Reason: It fails to highlight the significance of the research, especially around clinical practices. But we do understand the broad topic that study is about two surgical treatments for popliteal cysts.
- Layness: 3
  - It is a mix of simple and complex sentences compared to the abstract. However, it contains significant medical jargon and complex sentences.
- Usefulness: 2
  - The summary provides limited information due to complex sentences and medical jargon
- Factuality: 3
  - Reason: Frequent typo mistakes but they don't alter the meaning. E.g, arroscopy, "Lindberg" instead of "Lindgren"

Model Y:

This study compared two surgical treatments for popliteal cysts in patients. One group had internal drainage surgery, and the other group had internal drainage plus cyst wall removal using special techniques. MRI scans were done to see if the cysts came back. The patients were followed up for about two years. The group that had cyst wall removal had longer surgeries and more complications. Both groups improved in knee function, but there were no significant differences between them at the end. The study concluded that adding cyst wall removal reduces cyst recurrence but increases surgery time and complications.

- Comprehensiveness: 4
  - Reason minor omission of the number of patients and group details.
- Layness: 5
- Usefulness: 5
- Factuality: 5

Abstract 2:

LncRNA-protein interactions play important roles in post-transcriptional gene regulation , poly-adenylation , splicing and translation . Identification of lncRNA-protein interactions helps to understand lncRNA-related activities . Existing computational methods utilize multiple lncRNA features or multiple protein features to predict lncRNA-protein interactions , but features are not available for all lncRNAs or proteins; most of existing methods are not capable of predicting interacting proteins (or lncRNAs) for new lncRNAs (or proteins) , which don't have known interactions . In this paper , we propose the sequence-based feature projection ensemble learning method , “SFPEL-LPI” , to predict lncRNA-protein interactions . First , SFPEL-LPI extracts lncRNA sequence-based features and protein sequence-based features . Second , SFPEL-LPI calculates multiple lncRNA-lncRNA similarities and protein-protein similarities by using lncRNA sequences , protein sequences and known lncRNA-protein interactions . Then , SFPEL-LPI combines multiple similarities and multiple features with a feature projection ensemble learning frame . In computational experiments , SFPEL-LPI accurately predicts lncRNA-protein associations and outperforms other state-of-the-art methods . More importantly , SFPEL-LPI can be applied to new lncRNAs (or proteins) . The case studies demonstrate that our method can find out novel lncRNA-protein interactions , which are confirmed by literature . Finally , we construct a user-friendly web server , available at http://www . bioinfotech . cn/SFPEL-LPI/

Model X:

Scientists have found that interactions between lncRNAs (long non-coding RNAs) and proteins are very important for controlling how genes work after they are made into RNA. These interactions help with processes like adding a tail to RNA (poly-adenylation), cutting and joining RNA pieces (splicing), and making proteins from RNA (translation). Understanding these interactions helps us learn more about what lncRNAs do in the cell. Many current computer methods try to predict which lncRNAs and proteins will interact by looking at various features of lncRNAs or proteins. However, these features are not always known for all lncRNAs or proteins, making it hard to predict new interactions. In this study, we introduce a new method called SFPEL-LPI, which stands for Sequence-based Feature Projection Ensemble Learning for lncRNA-Protein Interactions. This method uses the sequences of lncRNAs and proteins to find features. It then calculates how similar different lncRNAs and proteins are to each other based on their sequences and known interactions. By combining these similarities and features, SFPEL-LPI can accurately predict which lncRNAs and proteins will interact. Our experiments show that SFPEL-LPI is better at predicting these interactions than other methods. Importantly, it can also predict interactions for new lncRNAs or proteins that have not been studied before. We tested our method and found new lncRNA-protein interactions that were confirmed by other scientific studies. We have construct a user-friendly web server, available at http://www.bioinfotech.cn/SFPEL-LPI/

- Comprehensiveness: 5
- Layness: 5
- Usefulness: 5
- Factuality: 5

Model Y:

LncRNA-protein interactions play important roles in post-transcriptional gene regulation, poly-adenylation, splicing and translation. In this paper, we propose the sequence-based feature projection ensemble learning method, “SFPEL-LPI”, to predict lncRNAs and proteins. First, we extract multiple lncRNA features and protein sequence- based features. Second, we calculate multiple similarities and protein-protein similarities by using lnc RNA sequences, protein sequences and known lnc RNAs. Finally, we construct a user-friendly web server, available at http://www. bioinfotech. Cn/SFPE.

- Comprehensiveness: 2
    - Reason: The plain text is partially complete. We get just get a very vague idea of the study
- Layness: 2
    - Reason: The plain text generation just omitted a few sentences from the original abstract.
- Usefulness: 1
    - Reason: Unable to understand the study due to medical jargon and fine-grained details than a high-level overview presented in the summary.
- Factuality: 2
    - Reason: Alters the meaning of parts of the study e.g, “SFPEL-LPI is used "to predict lncRNAs and proteins” but it is used to predict the interaction between them.