**NLPCC2024 Shared Task 6**

**Competition Proposal for Scientific Literature Survey Generation**

**Guideline**

Kexin Technology, Deakin University

April 15th, 2024

### **Competition background**

Doing literature survey is a necessary part of scientific research. However, it takes more and more time for researchers to do this, since the amount of scientific publications is vast and the publishing speed is still increasing rapidly. AI-generated literature survey can help researchers to save time. It becomes more realistic for researchers to adopt AI-generated literature survey with the power of large language model (LLM).

In this competition, we will further investigate the possibility and evaluate the usability of AI-generated literature survey. Given a list of topic-related references, the task aims to develop models for long form literature survey generation. We provide a training set which contains randomly crawled arXiv survey papers together with their references. For each survey paper, participants are expected to provide a corresponding AI-generated survey by developing retrieval augmented generation models based on the content and references in the original survey paper. Both automatic and human evaluation will be conducted to measure the generated survey outline and content quality. There is no limit for the use of large language models (e.g. LLaMA2, GPT4, etc.), however, reverse engineering is not allowed.

### **Competition task**

To develop retrieval augmented text generation models for scientific literature survey generation, based on the given metadata (title, subject, abstract, references) of scientific research literature survey papers and corresponding survey text in different scientific fields.

### **Deliverables**

Model development code, report, and generated surveys.

### **Training data**

We provide around 500 English scientific research survey papers, of which 400 can be used as training set and 100 can be used as verification set. Each sample data contains the title, article\_id, subject, abstract, reference, reference\_content of the article and survey content. (Note: data other than the training set cannot be used in the process of model development).

The sample of training data is provided as JSON files which are displayed as Fig 1. “reference\_content” contains title and abstract for part of references.

### **Testing data**

200 survey papers are provided as testing set. The testing datasets will be provided as a JSON file in format of {“subject”: “”, “reference”: [...]}

### **Evaluation criteria**

Automatic evaluation is mainly based on ROUGE, Soft Heading Recall1, LLM and human evaluation will also be conducted to measure the survey outline and content quality. Each method will be accounted for 25% weight.

1. We utilize ROUGE-1/2/L provided by Google2 to evaluate the content of generated survey.
2. We utilize Soft Heading Recall for evaluating the structure of generated survey.

represents all the chapter titles in a generated survey, the bge-large-en-v1.53 model will be using for text embedding.

1. We utilize LLM and human experts to evaluate the following aspects:

* Fluent language with clear expression;
* Logical article structure;
* Ample, reliable, and accurate citations;
* Consistency of content with the theme, staying on-topic;
* Broad analytical scope.

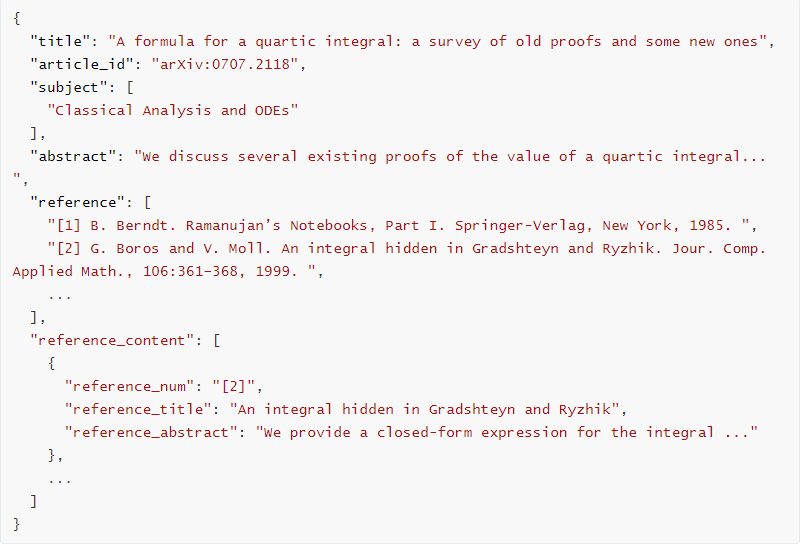


Fig 1. The format of training data sample

**Organizer:** Kexin Technology, Deakin University

**Contact:** Yangjie Tian ([rd\_kexin@cnpiec.com.cn](mailto:rd_kexin@cnpiec.com.cn))