NTCIR-18

Core Task: Automatic Evaluation of LLMs (AEOLLM)

Junjie Chen, Zhumin Chu, Haitao Li, Qingyao Ai, Yiqun Liu*

Department of Computer Science & Tech., Tsinghua University

chenjj826@gmail.com, aiqy@tsinghua.edu.cn

NTCIR 清華大學 Tsinghua University

Motivation

- As LLMs grow popular in both academia and industry, how to effectively evaluate the capacity of LLMs becomes an increasingly critical but still challenging issue.
- The existing LLM evaluation methods could be categorized into two groups:

manual evaluation

the most effective and reliable, but **high cost**

automatic evaluation

reduce human involvement, more objective

more promising

==> NCTIR-18 Automatic Evaluation of LLMs (AEOLLM)

NTCIR 清華大學 Tsinghua University

Methodology

- AEOLLM has two main characteristics:
  1. concentrates on generative tasks
  2. encourages reference-free evaluation methods.

Methodology

- AEOLLM has two main characteristics:
  1. concentrates on generative tasks
  2. encourages reference-free evaluation methods.

multiple-choice-format questions: easy to process, but this format differs from the real-world practical questions, which usually don’t have definite answers.

generative tasks: evaluate the capacity of automatic evaluation methods in assessing open-ended responses.

Methodology

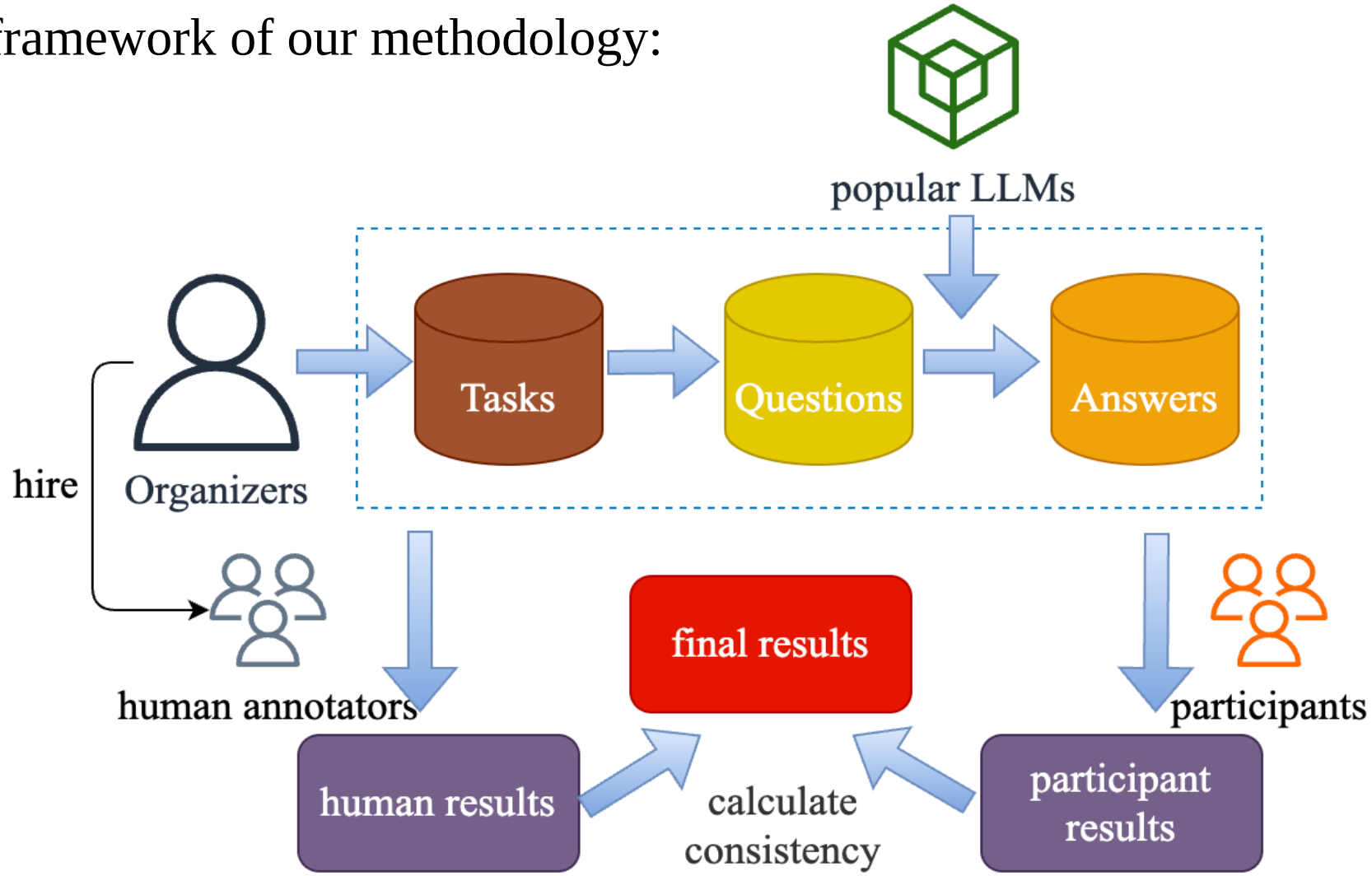
- AEOLLM has two main characteristics:
  1. concentrates on generative tasks
  2. encourages reference-free evaluation methods.

reference-based metrics (such as Rouge and BLEU): widely used, but cannot accurately reflect the quality of the results.

the gold reference can be trained rapidly by LLMs and then become useless.

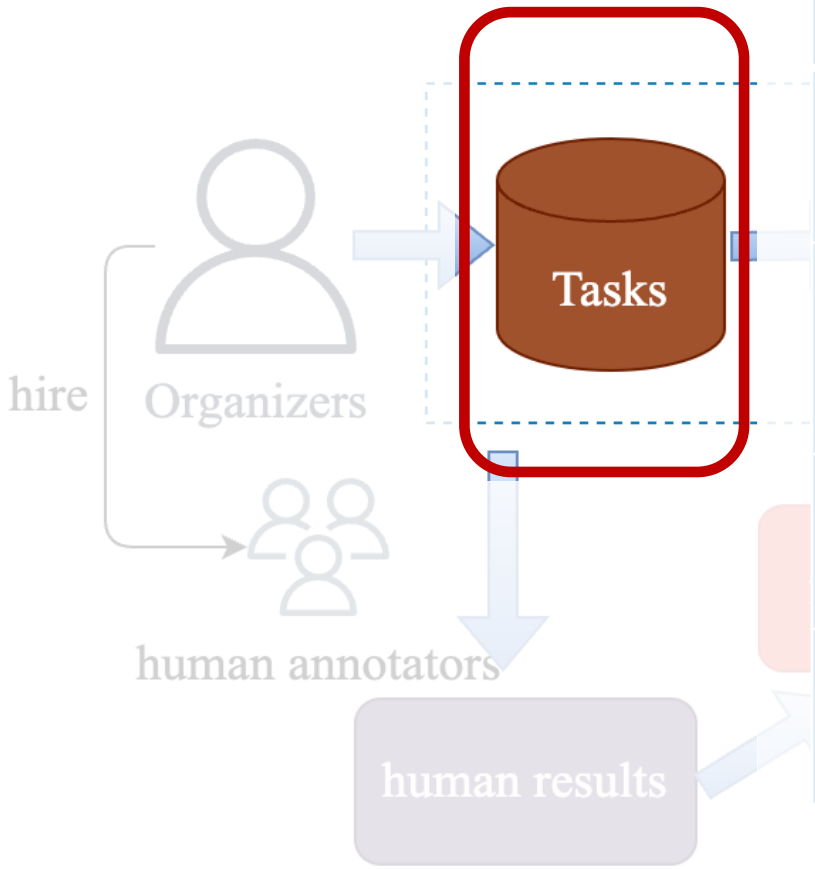
Methodology

- The framework of our methodology:



Methodology

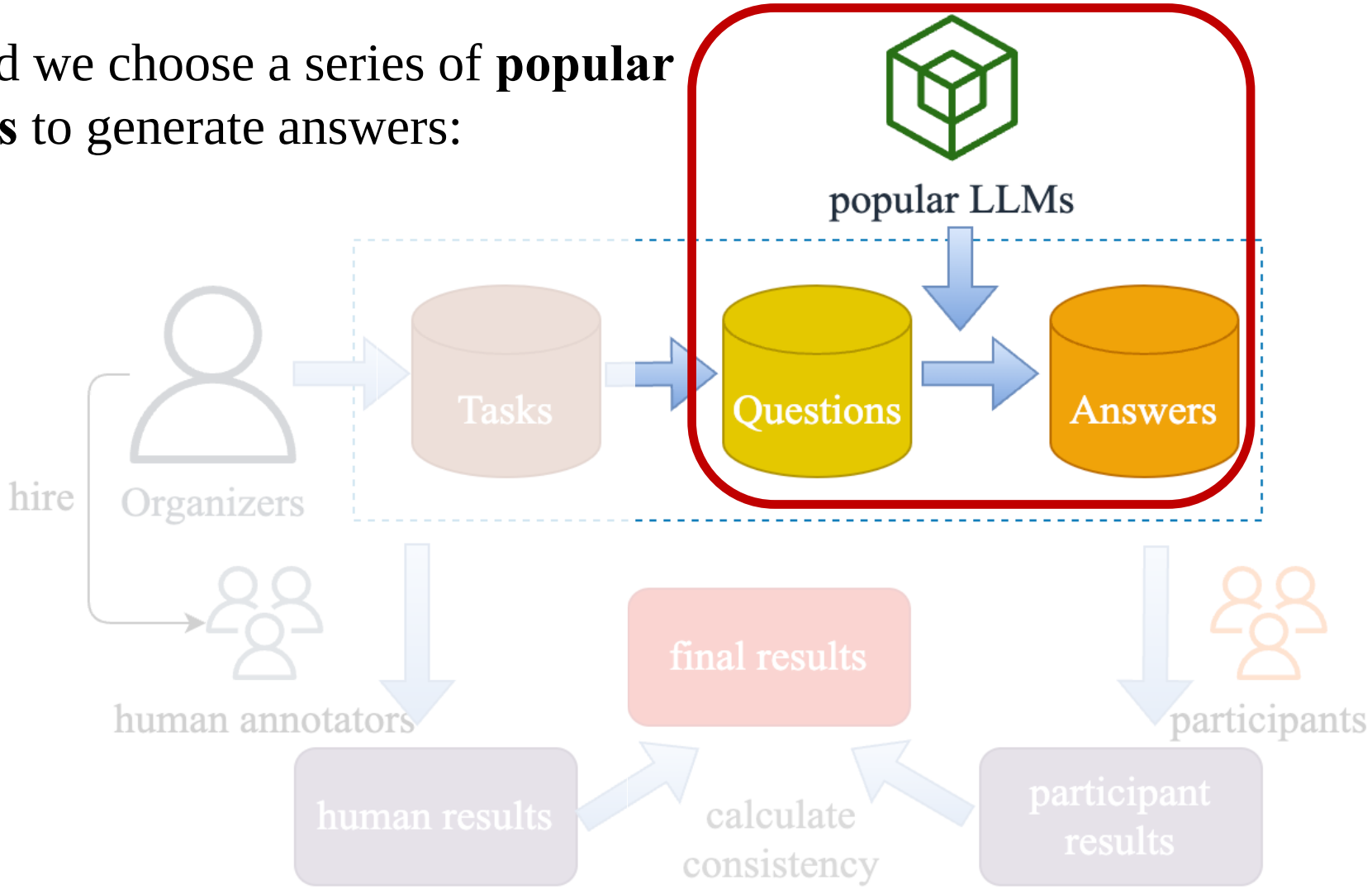
- First, we choose **four subtasks** as shown in the table below:



Task	Description
Summary Generation (SG)	write a summary for the specified text
Non-Factoid QA (NFQA)	construct long-form answers to open-ended non factoid questions
Text Expansion (TE)	generate stories related to the given theme
Dialogue Generation (DG)	generate human-like responses to daily topics

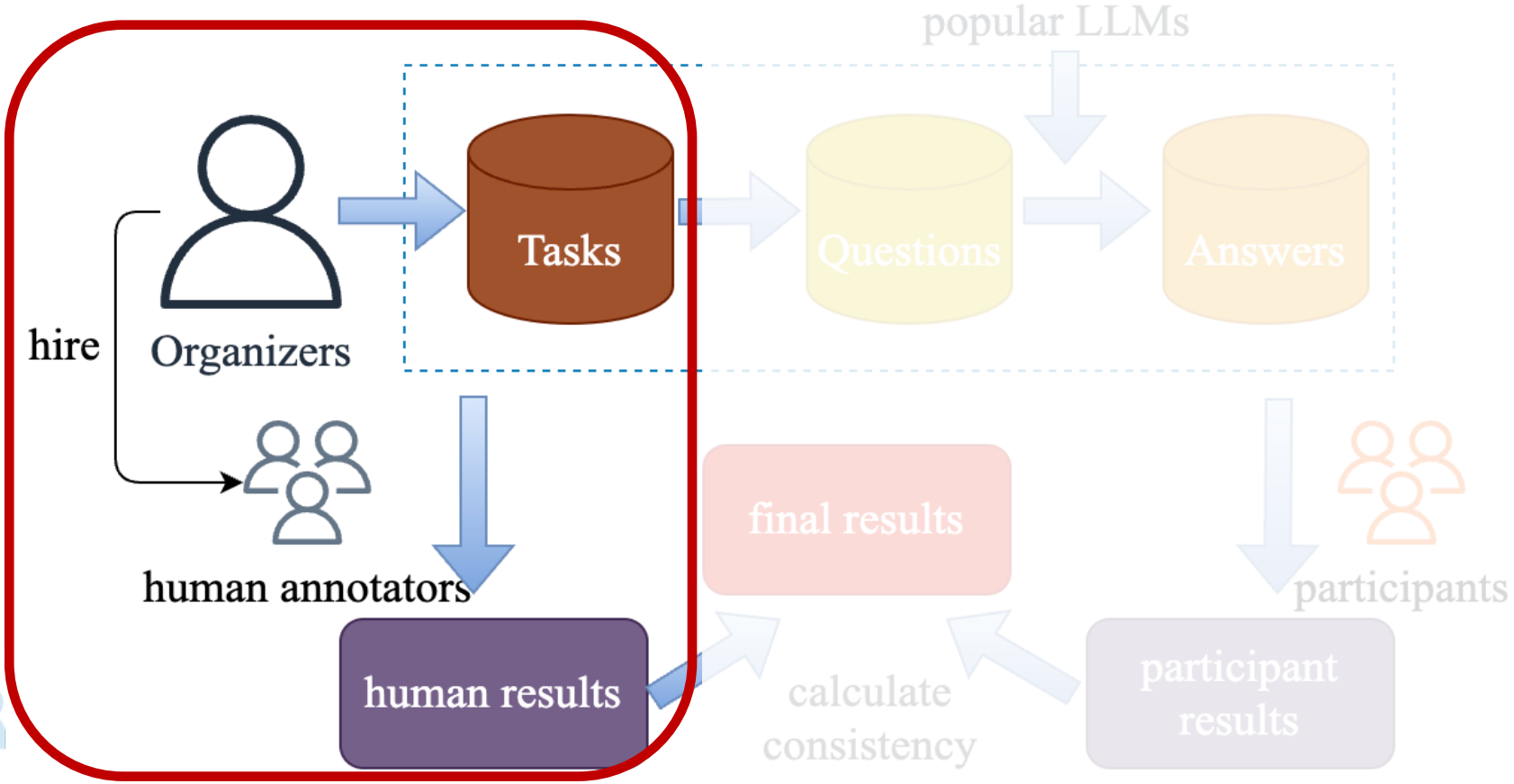
Methodology

- Second we choose a series of **popular LLMs** to generate answers:



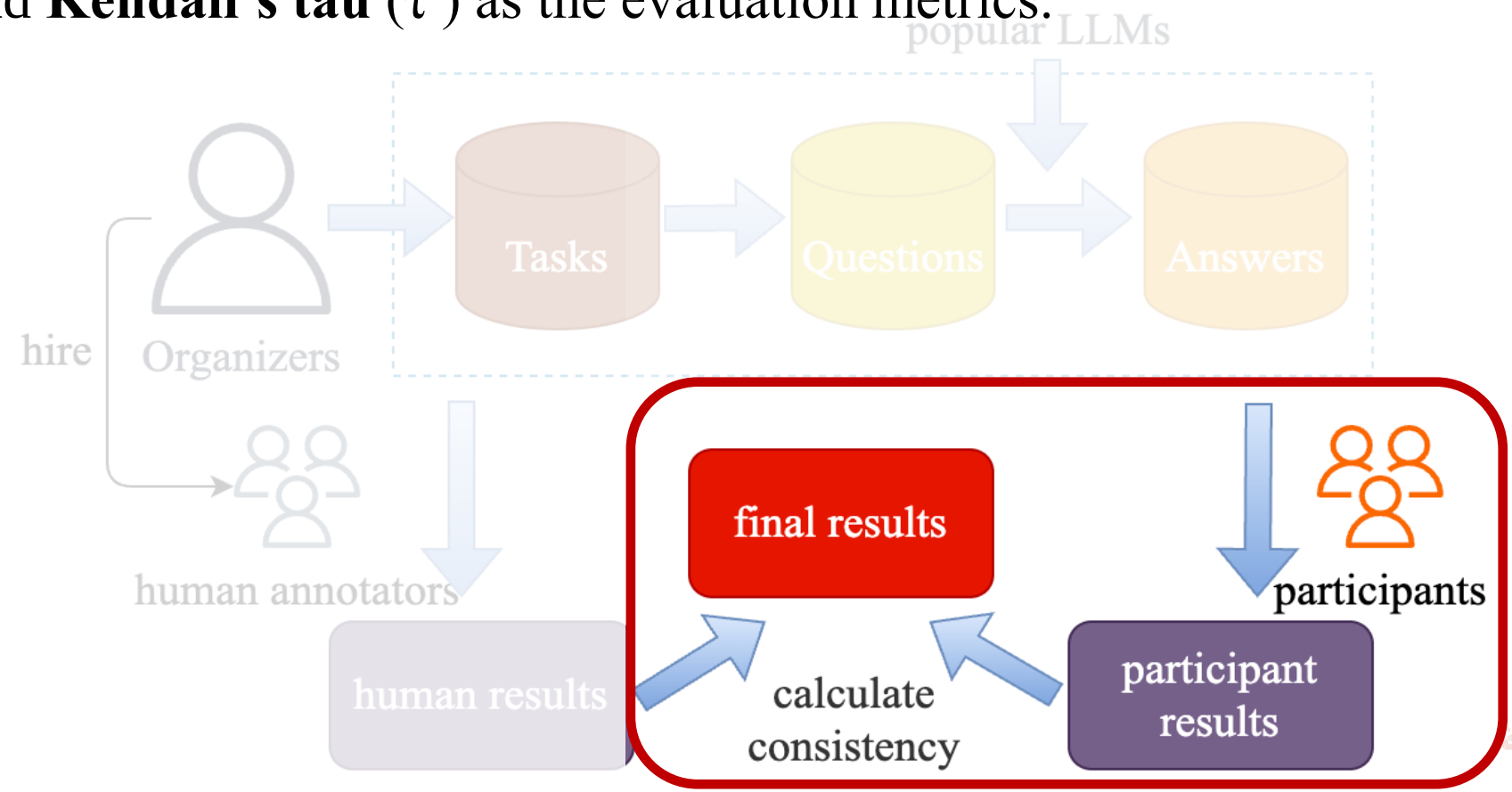
Methodology

- Third we manually annotate the answer sets for each question, which will be used as **gold standards** for evaluating the performance of different evaluation methods.



Methodology

- Last, we will collect evaluation results from participants and calculate consistency with manually annotated results. We will use **Spearman correlation coefficient** (S) and **Kendall's tau** (τ) as the evaluation metrics.



Dataset & Resources

- **Summary Generation (SG): Xsum** (https://huggingface.co/datasets/EdinburghNLP/xsum)
  - A real-world single-document news summary dataset collected from online articles by the British Broadcasting Corporation (BBC) and contains over 220 thousand news documents.
- **Non-Factoid QA (NFQA): NF_CATS** (https://github.com/Lurunchik/NF-CATS)
  - A dataset contains examples of 12k natural questions divided into eight categories and doesn't have gold reference.
- **Text Expansion (TE): WritingPrompts** (https://huggingface.co/datasets/euclaise/writingprompts)
  - A large dataset of 300K human-written stories paired with writing prompts from an online forum.
- **Dialogue Generation (DG): DailyDialog** (https://huggingface.co/datasets/daily_dialog)
  - A high-quality dataset of 13k multi-turn dialogues. The language is human-written and less noisy.

Organizers

✉Yiqun Liu [yiqunliu@tsinghua.edu.cn] (Tsinghua University)

✉Qingyao Ai [aiqy@tsinghua.edu.cn] (Tsinghua University)

✉Junjie Chen [chenjj826@gmail.com] (Tsinghua University)

✉Zhumin Chu [chuzm19@mails.tsinghua.edu.cn] (Tsinghua University)

✉Haitao Li [liht22@mails.tsinghua.edu.cn] (Tsinghua University)

NTCIR

清華大學 Tsinghua University

Schedule

- March 2024: Kickoff Event
- May 2024: Dataset release*
- Jun-Dec 2024: Dry run*
- Sep 2024-Feb 2025: Formal run*
- Feb 1, 2025: Evaluation results return
- Feb 1, 2025: Task overview release (draft)
- Mar 1, 2025: Submission due of participant papers (draft)
- May 1, 2025: Camera-ready participant paper due
- Jun 10-13 2025: NTCIR-18 Conference
- (* indicates that the schedule can be different for different tasks)

NTCIR 清華大學 Tsinghua University

References

[1] Rohaid Ali, Oliver Young Tang, Ian David Connolly, Patricia L Zadnik Sullivan, John H Shin, Jared S Fridley, Wael F Asaad, Deus Cielo, Adetokunbo A Oyelese, Curtis E Doberstein, et al. Performance of chatgpt and gpt-4 on neurosurgery written board examinations. medRxiv, pages 2023–03, 2023.

[2] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. Gpteval: A survey on assessments of chatgpt and gpt-4. arXiv preprint arXiv:2308.12488, 2023.

[3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xi_x0002_aoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 2023.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

References

[7] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the sum_x0002_mary! topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv:1808.08745, 2018.

[8] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. A non_x0002_factoid question-answering taxonomy. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1196–1207, 2022.

[9] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. arXiv preprint arXiv:1805.04833, 2018.

[10] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957, 2017.

[11] Ann Lehman, Norm O'Rourke, Larry Hatcher, and Edward Stepanski. JMP for basic univariate and multivariate statistics: methods for researchers and social scientists. Sas Institute, 2013.

[12] Maurice G Kendall. A new measure of rank correlation. Biometrika, 30(1/2):81–93, 1938.

Thank you!

NTCIR

清華大學 Tsinghua University