NTCIR-19

Core Task: Automatic Evaluation of LLMs 2 (AEOLLM 2)

Junjie Chen, Haitao Li, Qingyao Ai, Yiqun Liu

Department of Computer Science & Tech., Tsinghua University

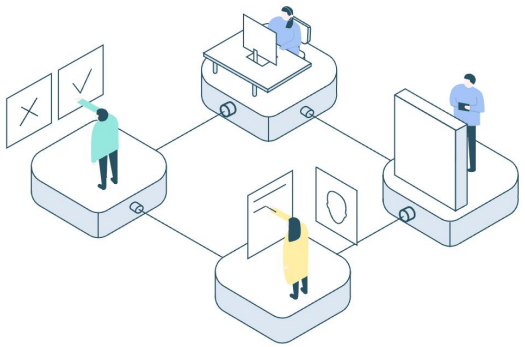
chenjj826@gmail.com, aiqy@tsinghua.edu.cn

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.
- Current evaluation methods for LLMs 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

Motivation

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

Motivation

- 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.

Motivation

- 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.

Motivation

- ***Challenge:*** Through AEOLLM, our findings revealed that current automatic evaluation methods of LLMs still remain limited in effectiveness when applied to long-form generation tasks.

Table 3: The results from the formal run on the reserved set. Baselines 1, 2, 3, and 4 correspond to direct prompting of ChatGLM3-6B, Baichuan2-13B, ChatGLM-Pro, and GPT-4o, respectively. The best result is highlighted in bold.

Team	Dialogue Generation			Text Expansion			Summary Generation			Non-Factoid QA			Overall		
	acc	τ	ρ	*acc*	τ	ρ	*acc*	τ	ρ	*acc*	τ	ρ	*acc*	τ	ρ
Baseline1	0.5583	0.3228	0.3495	0.5029	0.1236	0.1293	0.5976	0.2589	0.2712	0.6445	0.3717	0.3948	0.5759	0.2692	0.2862
Baseline2	0.5518	0.1647	0.1750	0.5021	0.0698	0.0725	0.6112	0.2693	0.2813	0.6044	0.2664	0.2791	0.5674	0.1926	0.2020
Baseline3	0.5924	0.2921	0.3135	0.5495	0.2831	0.2985	0.7079	0.4390	0.4575	0.6975	0.4682	0.4990	0.6368	0.3706	0.3921
Baseline4	0.6595	0.4423	0.4797	0.5543	**0.3963**	**0.4248**	0.7029	0.3886	0.4180	0.7441	**0.4896**	**0.5281**	0.6652	0.4292	0.4627
KNUIR	0.6778	0.4404	0.4717	0.5512	0.3141	0.3430	0.7375	0.4524	0.4914	0.6951	0.4102	0.4297	0.6654	0.4043	0.4340
ISLab	/	/	/	0.5241	0.3609	0.4035	**0.7658**	0.5117	0.5632	/	/	/	/	/	/
UCLWI	**0.7756**	**0.5798**	**0.6426**	0.5266	0.3482	0.3815	0.7273	**0.5432**	**0.5763**	0.6853	0.4105	0.4291	0.6787	**0.4704**	**0.5074**
PanguIR	0.7444	0.5611	0.6091	**0.5581**	0.3432	0.3775	0.7479	0.5097	0.5520	**0.7528**	0.4175	0.4534	**0.7008**	0.4579	0.4980

NTCIR

清華大學 Tsinghua University

Motivation

- ***Growing trend:*** Use LLMs to generate **Deep Research Reports:**

in-depth research | **exceed 1500 tokens** | **factual accuracy** | **structured organization** | **thematic coherence**

Table 1: Statistics of benchmark datasets.

Dataset	Total	Average Length (tokens)			Label Count			Type	Type Count
		Prompt	Answer A	Answer B	First	Second	Tie		
RewardBench [11]	2985	439	165	163	1490	1495	0	Chat Chat_Hard Safety Reasoning	358 456 739 1432
MTBench [12]	2396	885	339	335	929	929	538	Turn1 Turn2	1204 1192
PreferenceBench [13]	1998	495	177	177	980	1018	0	-	-

Motivation

- ***Challenge:*** Through AEOLLM, our findings revealed that current automatic evaluation methods of LLMs still remain limited in effectiveness when applied to long-form generation tasks.

- ***Growing trend:*** Use LLMs to generate **Deep Research Reports:**

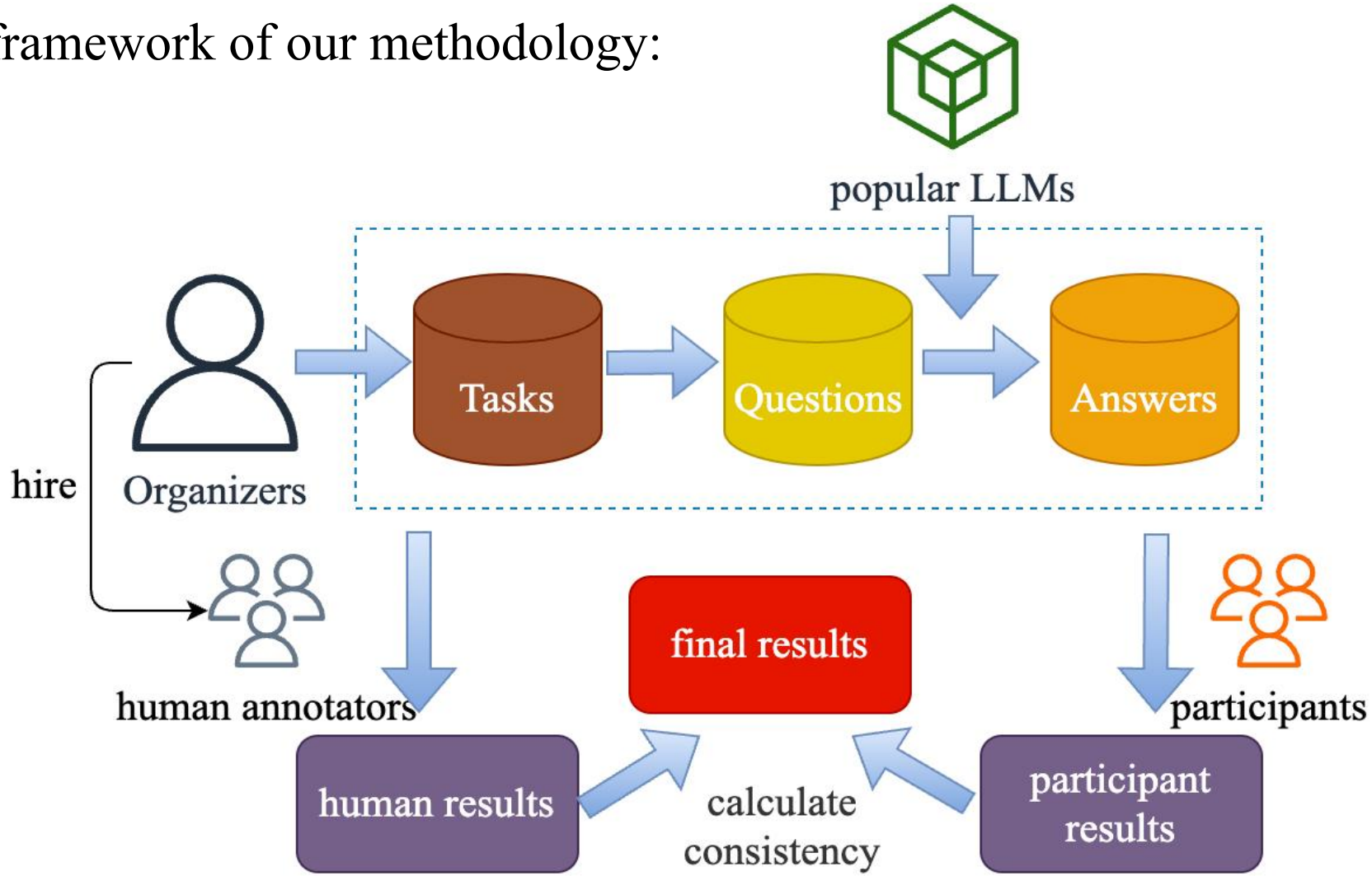
in-depth research | **exceed 1500 tokens** | **factual accuracy** | **structured organization** | **thematic coherence**

==> NCTIR-19 Automatic Evaluation of LLMs 2 (AEOLLM 2)

- focuses on the automated evaluation of long-form texts

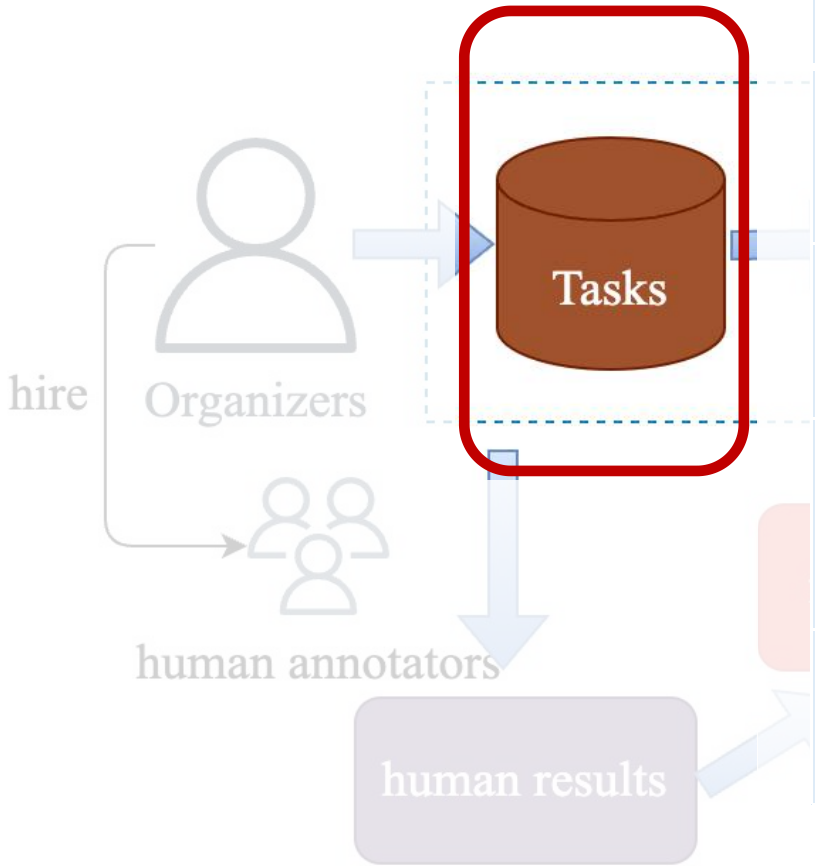
Methodology

- The framework of our methodology:



Methodology

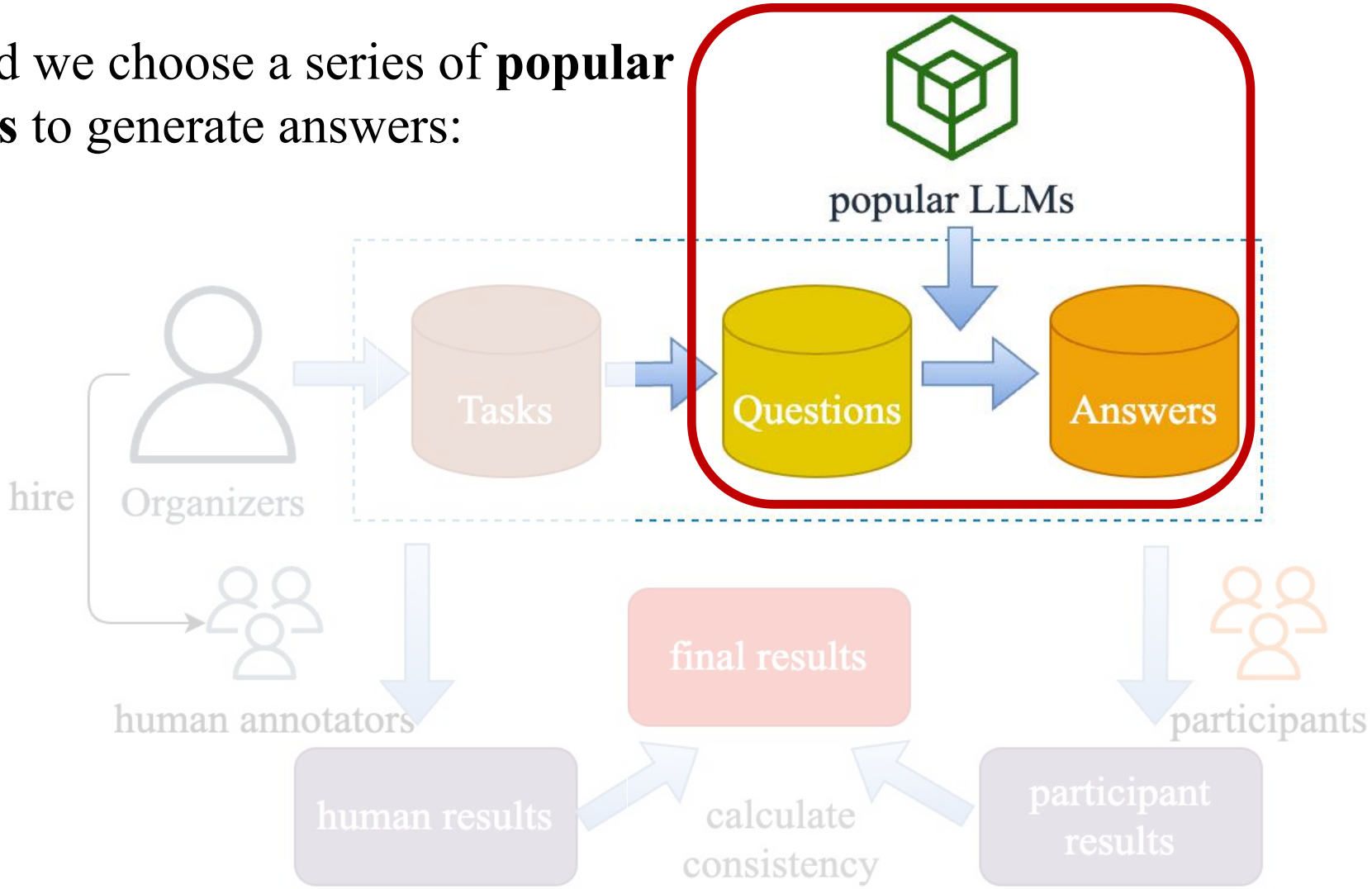
- First, we choose **multiple types of subtasks** as shown in the table below:



Task	Description
Text Expansion (TE)	Generate **long-form stories** related to the given theme
Survey Generation (SG)	Generate an **academic survey** on a given topic
Report Generation (RG)	Collect relevant information for a given question and generate an **in-depth report**
Legal Case Document Generation (LCDG)	Generate a complete multi-stage **legal case document**

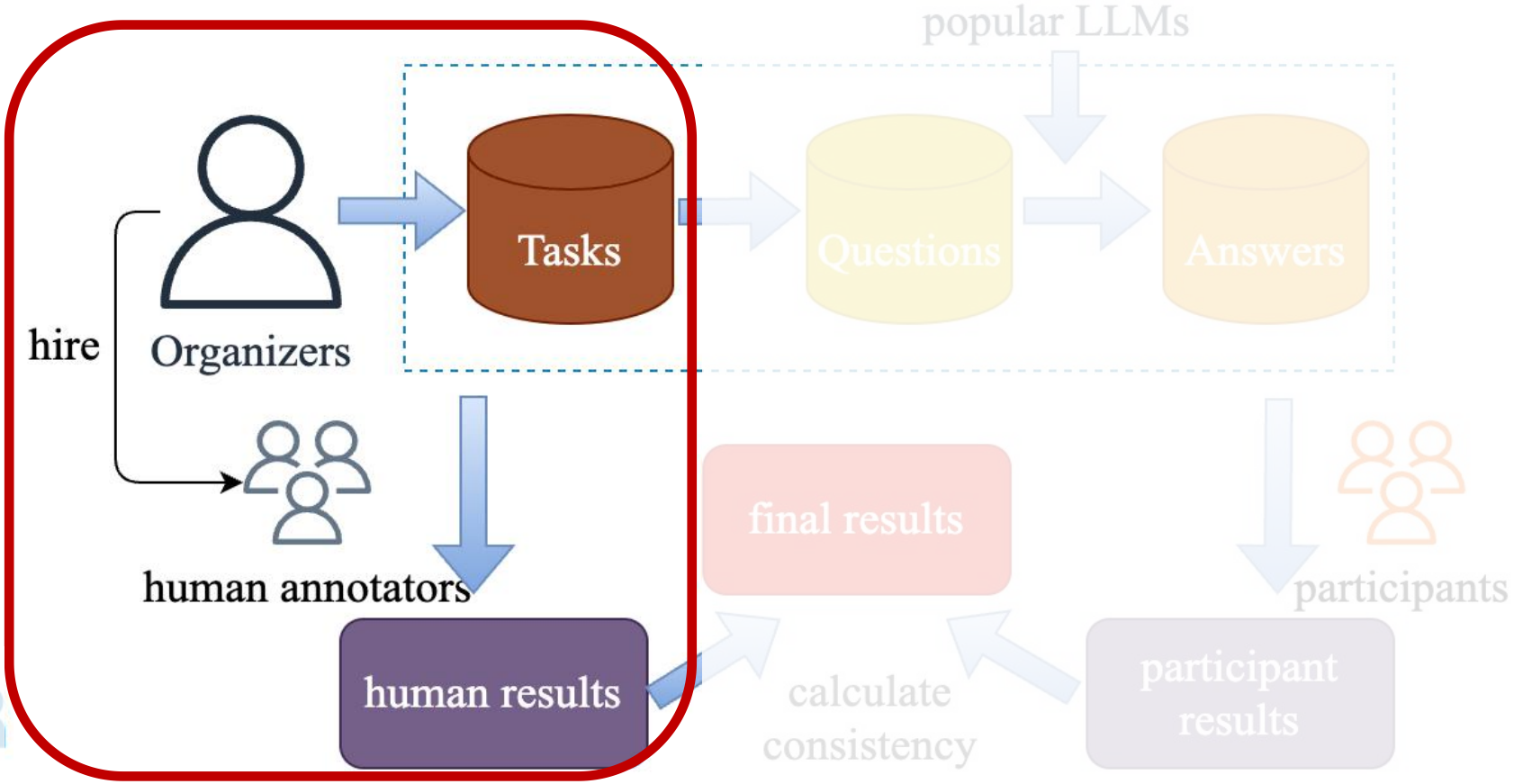
Methodology

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



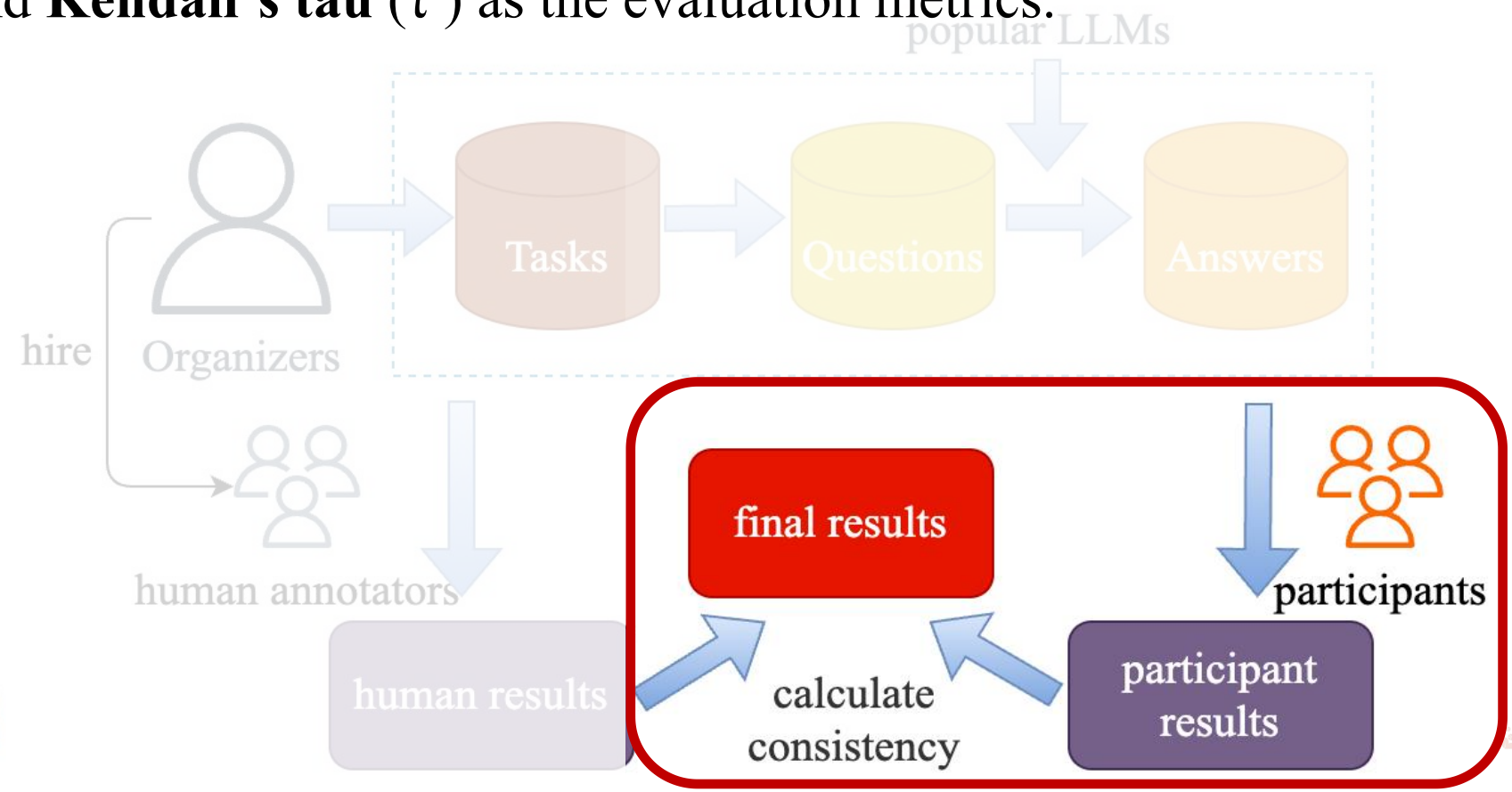
Methodology

- 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

- **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.
- **Survey Generation (SG): SurGE** (https://github.com/oneal2000/SurGE)
  - A dataset for end-to-end scientific survey generation in the computer science domain.
- **Report Generation (RG): DeepResearch Bench** (https://github.com/Ayanami0730/deep_research_bench)
  - A dataset consists of 100 PhD-level research tasks, each meticulously crafted by domain experts across 22 distinct fields
- **Legal Case Documents Generation (LCDG): CaseGen** (https://github.com/CSHaitao/CaseGen)
  - A dataset includes 500 real-world legal case instances, each structured into seven sections: Prosecution, Defense, Evidence, Events, Facts, Reasoning, and Judgment.

Organizers

- ✉ Junjie Chen [chenjj826@gmail.com] (Tsinghua University)
- ✉ Haitao Li [liht22@mails.tsinghua.edu.cn] (Tsinghua University)
- ✉ Yiqun Liu [yiqunliu@tsinghua.edu.cn] (Tsinghua University)
- ✉ Qingyao Ai [aiqy@tsinghua.edu.cn] (Tsinghua University)

NTCIR

清華大學 Tsinghua University

Schedule

- Sep 3, 2025: Kickoff Event
- January 2026: Dataset release*
- January-June 2026: Dry run*
- March-July 2026: Formal run*
- August 1, 2026: Evaluation results return
- August 1, 2026: Task overview release (draft)
- September 1, 2026: Submission due of participant papers (draft)
- November 1, 2026: Camera-ready participant paper due
- December 8-10, 2026: NTCIR-19 Conference
- (* indicates that the schedule can be different for different tasks)

NTCIR 清華大學 Tsinghua University

Thank you!

NTCIR

清華大學 Tsinghua University