Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling

AAAI 2024

Authors: Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, Yuesheng Zhu

Presenter: Jie Ruan

2024.2

CONTENTS

PART 01

Background

Background

北京大學 PEKING UNIVERSITY

Why is Reliable AI Generated Content Evaluation Important?

- Evaluation methods/criteria serve as the 'lighthouse' guiding the development of NLG technology:
  - Used to assess the performance of models/systems
  - Act as parameter tuning objectives
  - Serve as optimization targets for models

Background

PEKING UNIVERSITY

How to Evaluate AI Generated Content?

Challenge： Similar content in text can often be expressed in various ways, and the same output of the NLG system may need to satisfy multiple goals in different aspects

Three Factors：

Reproducibility

- Consistent results for multiple evaluations under the same setup (hardware, software, personnel, environment, etc.)
- Consistent results for multiple evaluations under different settings

Fairness

- Objectively reflect the quality of the generated text
- Fair comparison of different models/systems

Cost-efficient

- Low evaluation cost and high efficiency

Background

How to Evaluate AI Generated Content?

Challenge： Similar content in text can often be expressed in various ways, and the same output of the NLG system may need to satisfy multiple goals in different aspects

Human Evaluation

- ◆ Gold-standard
- ◆ Costly
- ◆ Low Reproducibility

Automatic Evaluation

- ◆ Unreliable
- ◆ Cheap
- ◆ High Reproducibility

PART 02

Introduction

Introduction

PEKING UNIVERSITY

Motivation

- To save labor and costs, researchers usually perform human evaluation on a small subset of data sampled from the whole dataset in practice.
- Problem of Random Sampling
  - Clustered Selection
  - Data Manipulation
  - Different selection subsets lead to **different inter-system rankings**

Experimental results from 137 real NLG evaluation setups on 44 human metrics across 16 datasets and 5 NLG tasks show **87.5% of datasets have different inter-system rankings across 5 times** of random sampling.

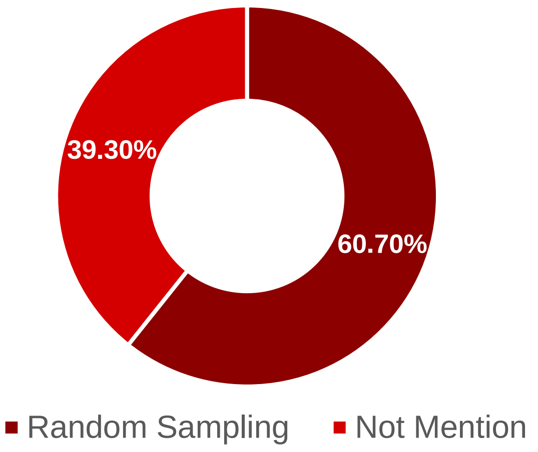


Fig. Survey on 1404 papers from top conferences about human evaluation sampling methods

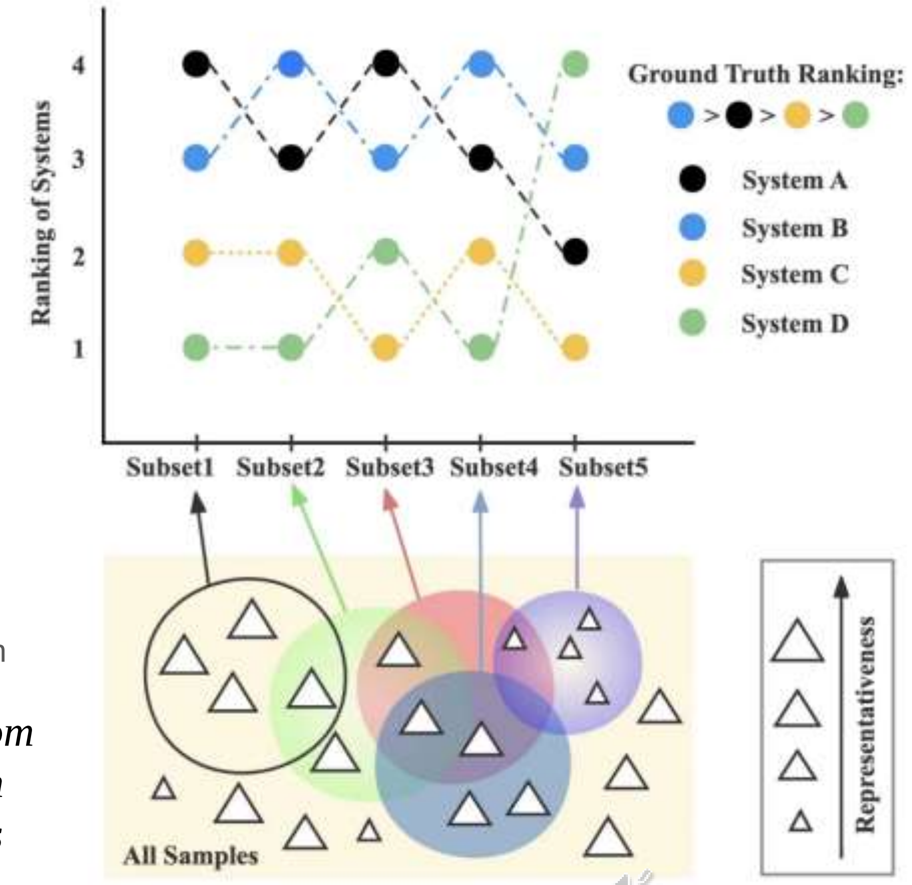


Fig. Random sampling is risky.

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

北京大學
PEKING UNIVERSITY

PART 02

Methodology

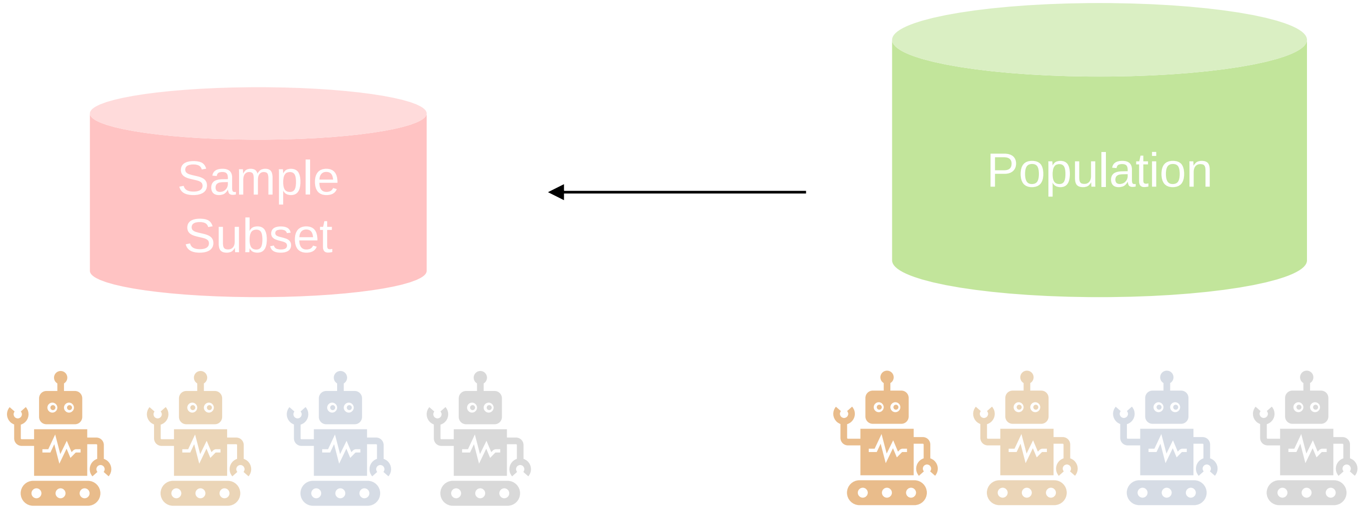
Methodology

Problem Statement

◆ The goal of sampling in human evaluation is to select part of the samples with the intention of **estimating the inter-system ranking of the whole sample population**. Ideally, the subset obtained by the sampling method should cover more **representative** samples of the population.

Sample for Evaluation

Ranking of
Generated AI Systems



[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

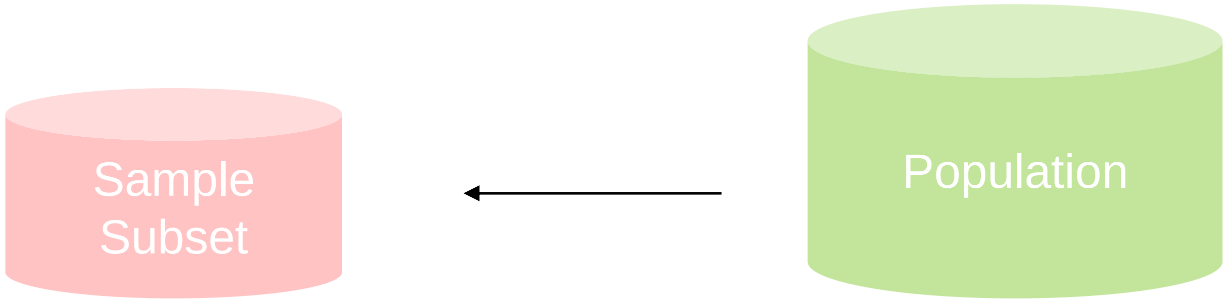
Methodology

PEKING UNIVERSITY

Sample Representativeness

- **Quality Diversity**: Evaluation on qualitatively diverse subsets of samples allows the system to better reflect the performance of all samples
- **Redundancy**: The degree of similarity or duplication among the generated outputs of samples

Sample for Evaluation



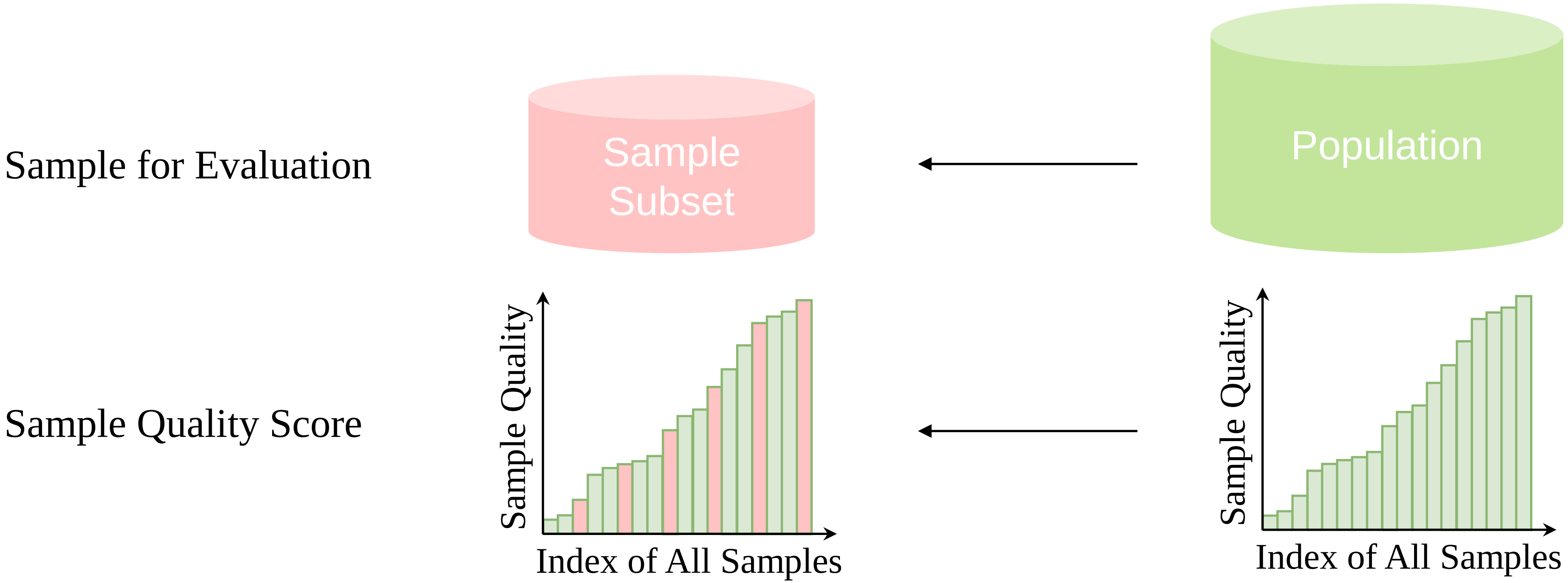
[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Methodology

PEKING UNIVERSITY

Sample Representativeness

- **Quality Diversity**: Evaluation on qualitatively diverse subsets of samples allows the system to better reflect the performance of all samples
- **Redundancy**: The degree of similarity or duplication among the generated outputs of samples



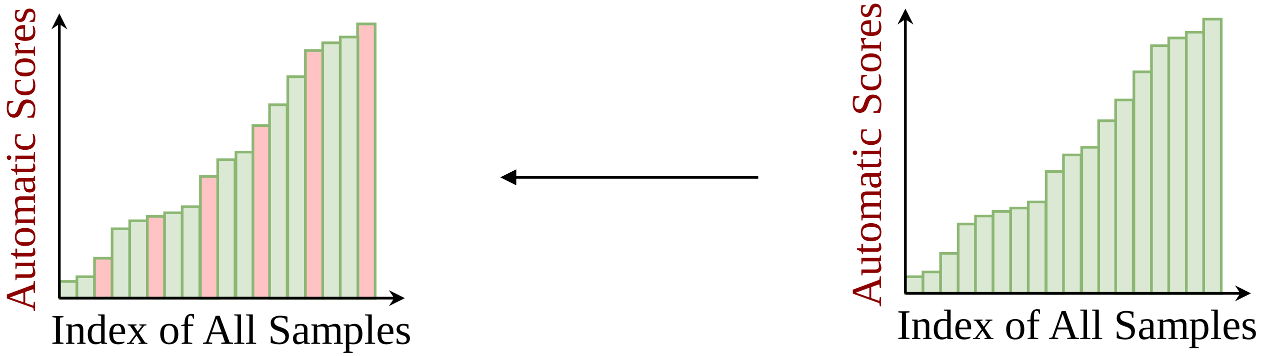
[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Methodology

How to Calculate Sample Quality Score? Utilizing Automatic Metrics

◆ As various automatic metrics can measure the characteristics of samples in different aspects and are easy to calculate with lower cost, we use scores of automatic metrics as features to predict the quality of samples.

Sample Quality Score



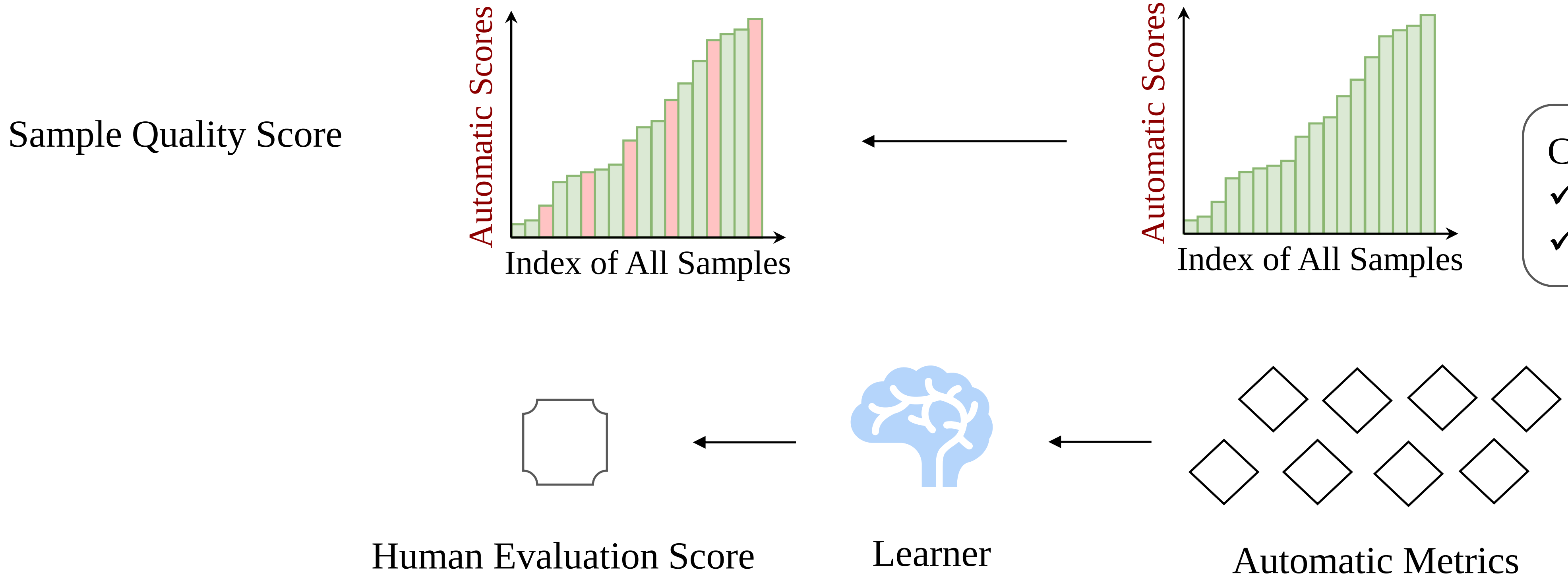
Comparison Method
✓ Single Metric (SM)
✓ 8 Metric (8M)

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Methodology

How to Calculate Sample Quality Score? Utilizing Automatic Metrics

- As various automatic metrics can measure the characteristics of samples in different aspects and are easy to calculate with lower cost, we use scores of automatic metrics as features to predict the quality of samples.



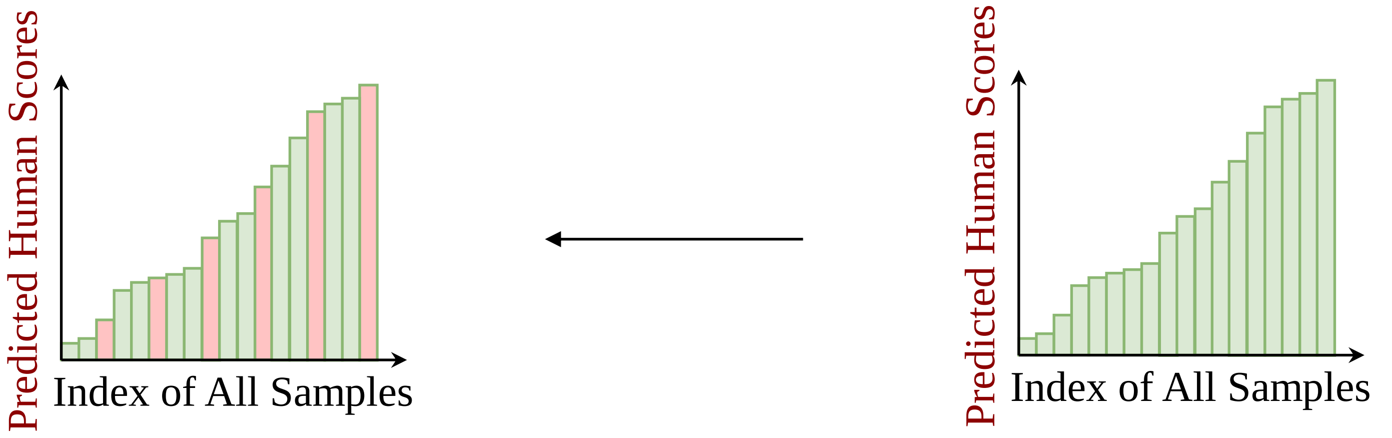
Comparison Method
✓ Single Metric (SM)
✓ 8 Metric (8M)

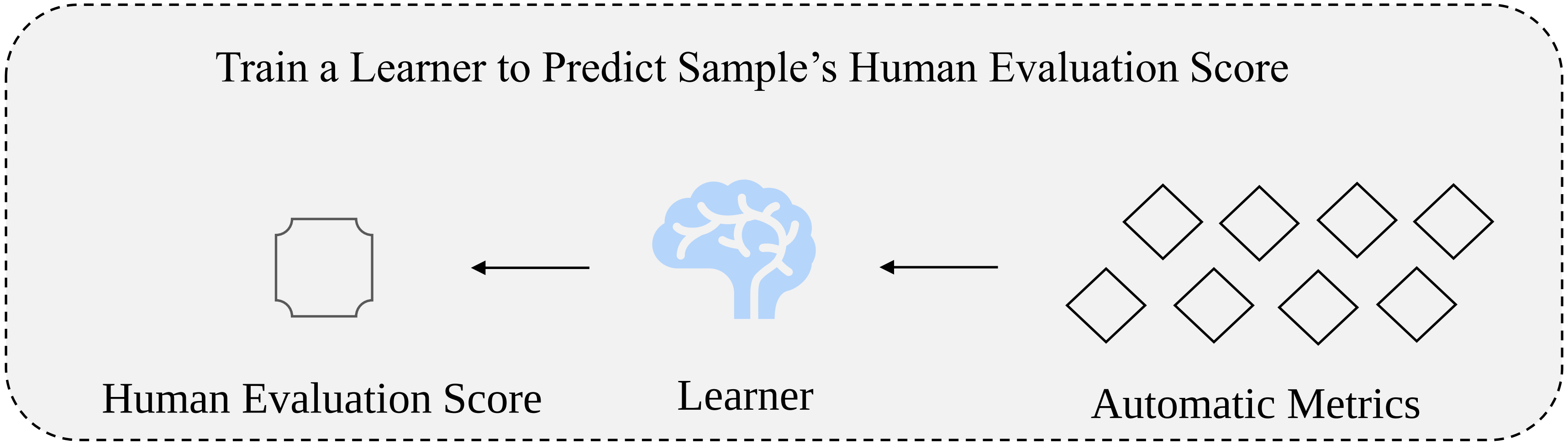
[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Methodology

PEKING UNIVERSITY

Learner

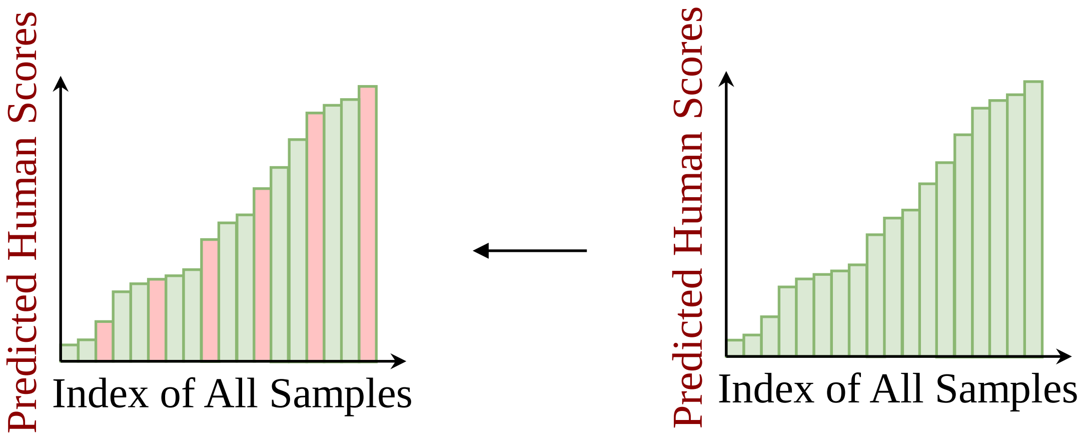




[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

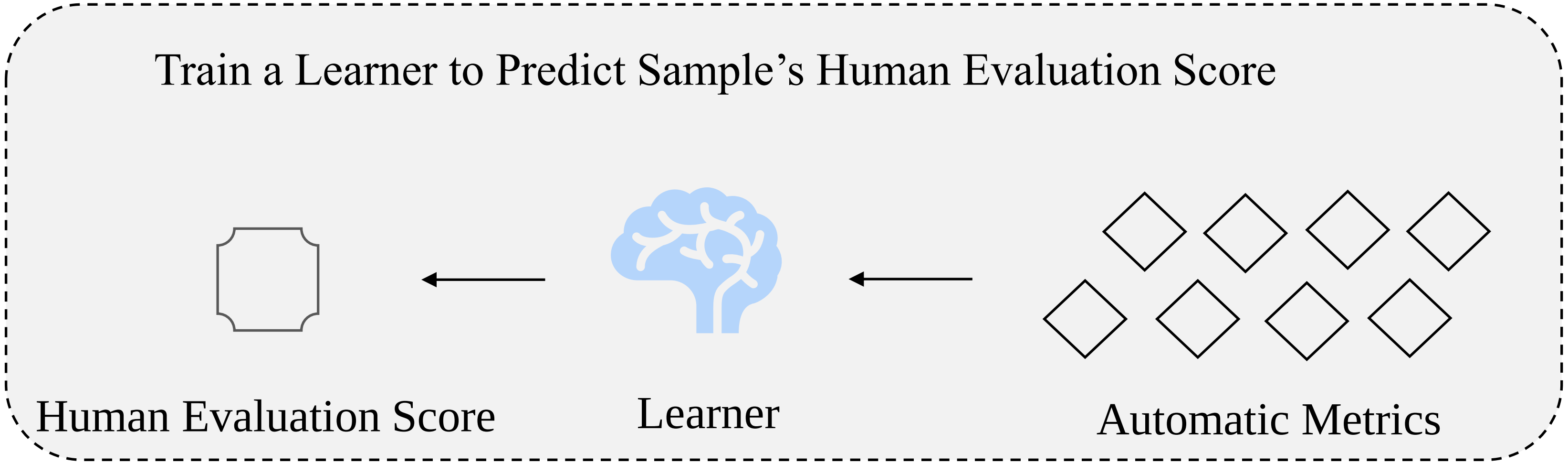
Methodology

Online Learning



Limited Training Data

◆ Online Learning



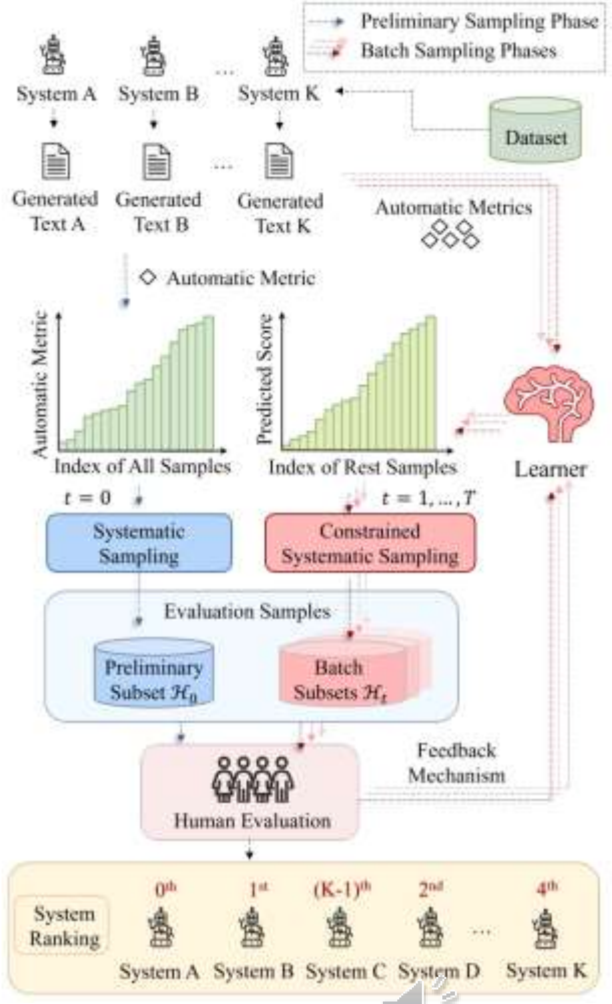


Fig. Constrained Active Sampling Framework

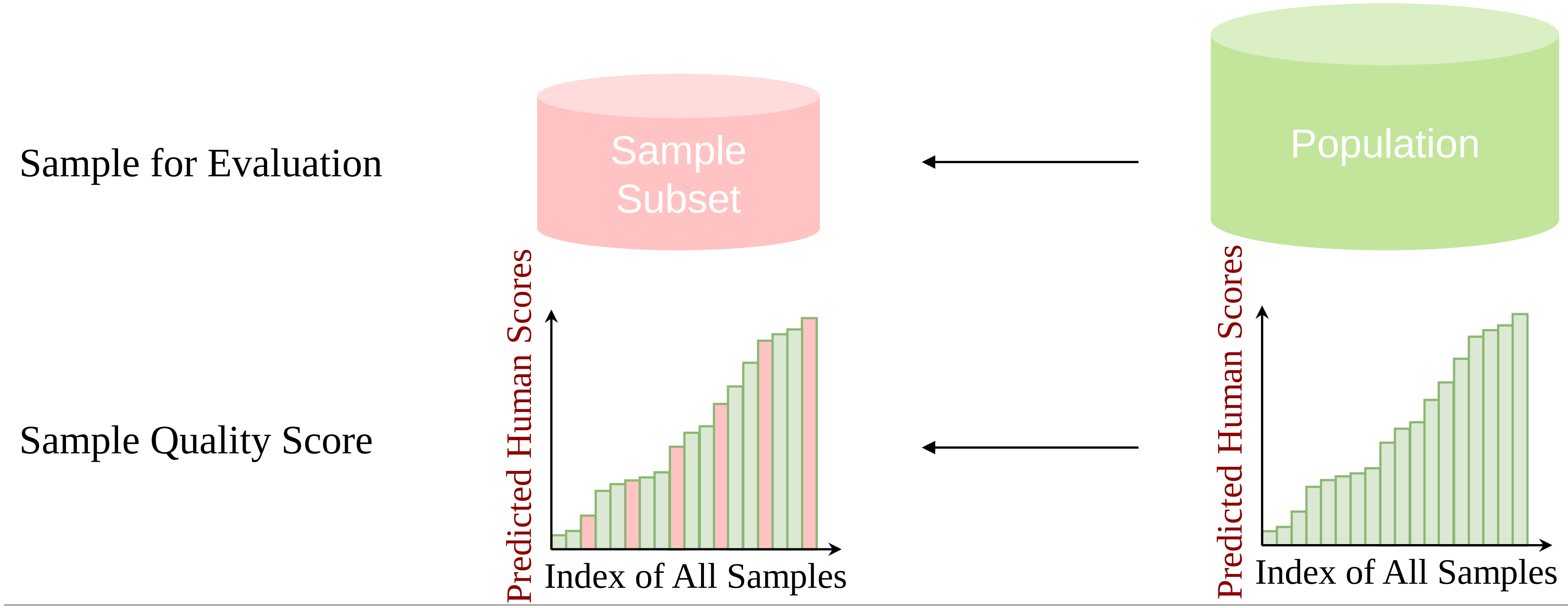
[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Methodology

Sample Representativeness

✓ **Quality Diversity**: Evaluation on qualitatively diverse subsets of samples allows the system to better reflect the performance of all samples

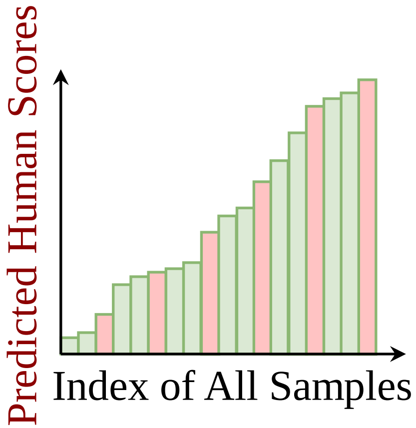
☐ **Redundancy**: The degree of similarity or duplication among the generated outputs of samples

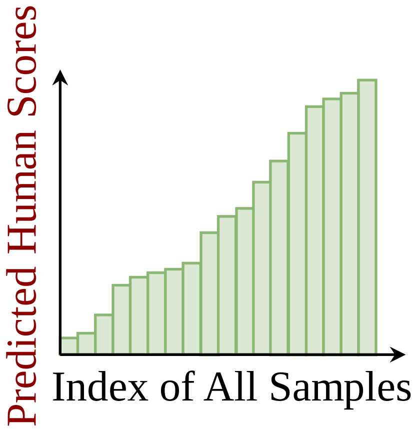


[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

北京大學 PEKING UNIVERSITY

Constrained Controller





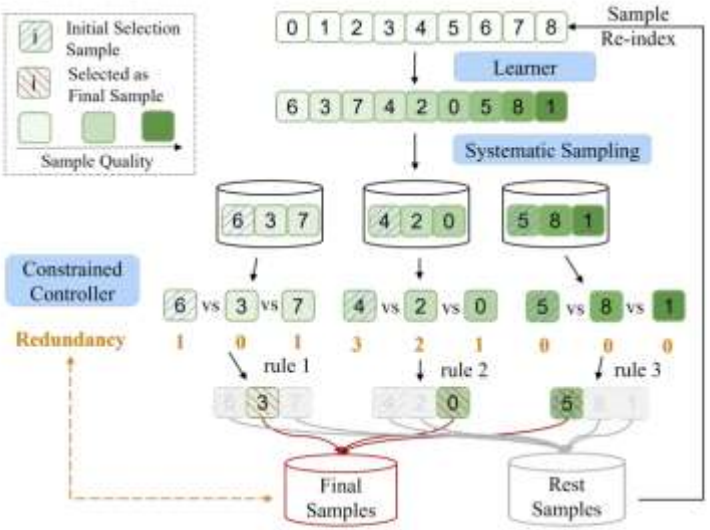


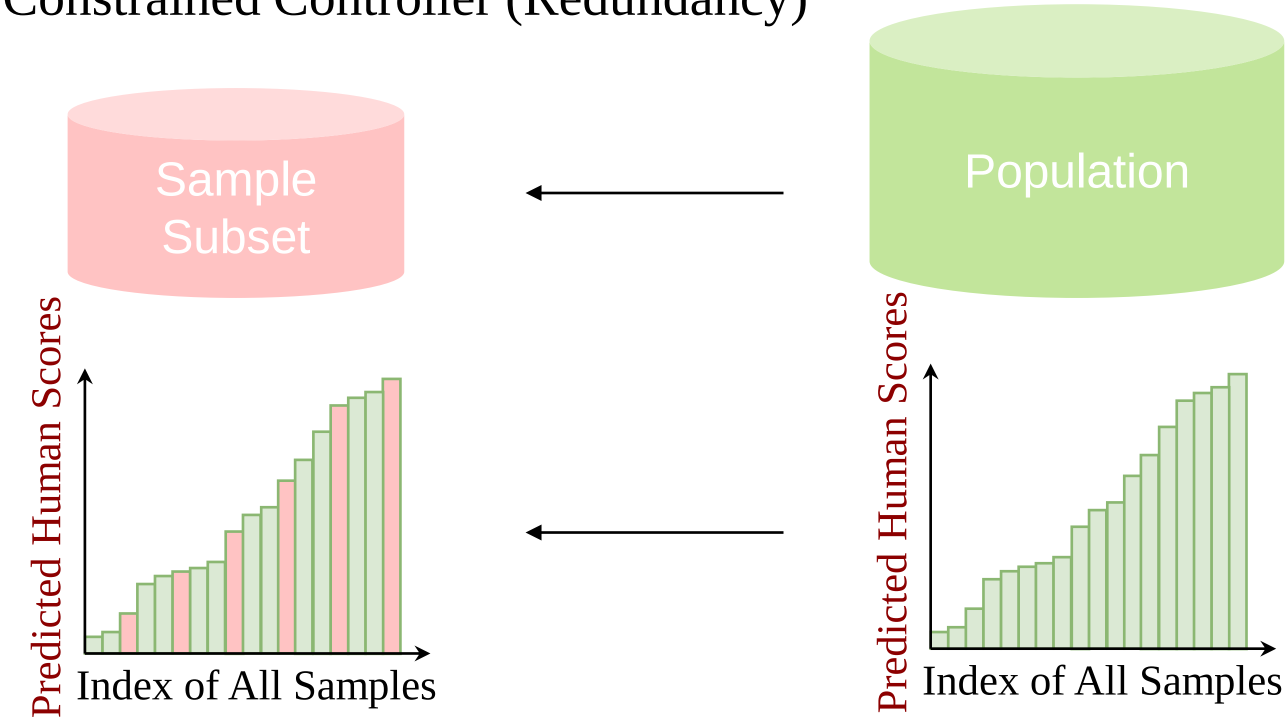
Fig. Constrained Controller

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Methodology

Constrained Active Sampling Framework

- Learner and Sample Quality
- Systematic Sampler
- Constrained Controller (Redundancy)



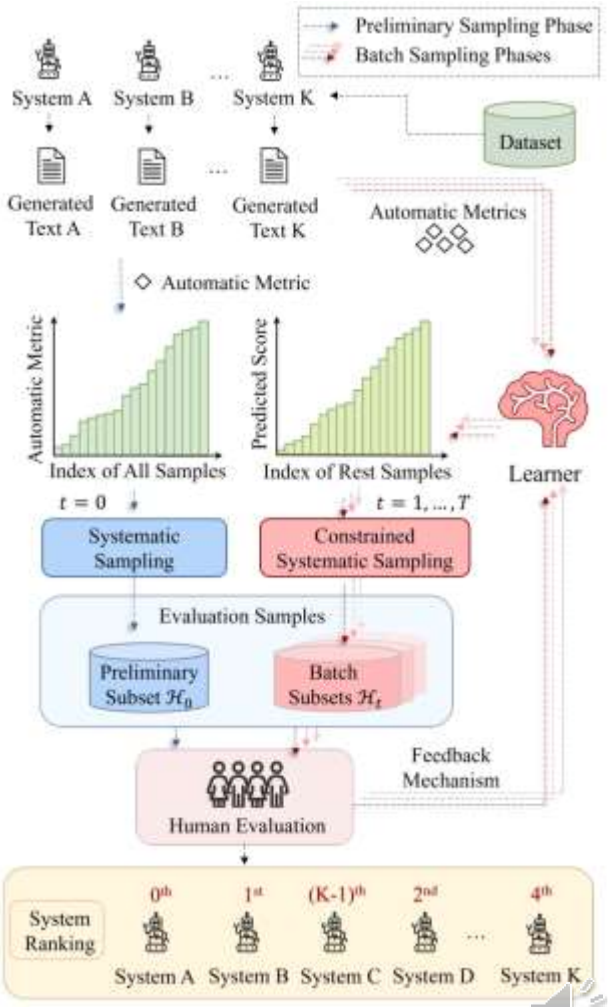


Fig. Constrained Active Sampling Framework

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

PART 04

Experiment

Experiment Setup

北京大學 PEKING UNIVERSITY

Tasks and Datasets

Summarization (SUM):

- SummEval (Fabbri et al. 2021)
- REALSumm (Bhandari et al. 2020)
- Newsroom (NeR18) (Grusky, Naaman, and Artzi 2018)
- DialSummEval (Gao and Wan 2022)
- OpenAI-axis1 (Stiennon et al. 2020; Volske et al. 2017)
- OpenAI-axis2
- OpenAI-CNN/DM1
- OpenAI-CNN/DM3

Machine Translation (MT):

- newstest2020 en-de
- newstest2020 cn-en
- newstest2021 cn-en (Freitag et al. 2021)

Dialogue Generation (DialoGen):

- Persona Chat (Mehri and Eskenazi 2020)

Story Generation (StoryGen):

- MANS-ROC (Guan et al. 2021)
- MANS-WP (Guan et al. 2021)

Multi-Modal Generation (MMGen):

- THUMB-MSCOCO (Kasai et al. 2022)
- VATEX-EVAL (Shi et al. 2022)

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Experiment Setup

PEKING UNIVERSITY

Evaluation Metrics

- **Kendall's Tau Correlation[1]**

Comparison of Methods

- **Random Sampling (Random)**
- **Heuristic Sampling (Heuristic):**
  - First sorts the samples according to the average sentence length of the sentences generated by all systems. Then, Heuristic randomly collects a small number of samples with extreme sentence length and a large number of samples with normal sentence length.
- **Eight Metric (8M)**
- **Single Metric (SM)**
- **Online Sampling (OL)**

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

Results and Analysis

PEKING UNIVERSITY 北京大學

Full Inter-System Ranking Accuracy

- Experiment results on 137 real NLG evaluation setups with 44 human evaluation metrics across 16 datasets and 5 NLG tasks demonstrate the proposed method ranks first and second on 95.45% of the human metrics with 0.83 overall inter-system ranking Kendall correlation.

Task	Dataset	HE Metric	Random 1	Random 2	Random 3	Random Mean	Heuristic 1	Heuristic 2	Heuristic 3	Heuristic Mean	8M	SM	OL	CASF (ours)
SUM	SummEval	coherence	0.8500	0.6500	0.3333	0.6111	0.7000	0.8167	0.9167	0.8111	0.4167	0.4167	0.8667	**0.9500**
		consistency	0.2500	0.4833	0.4333	0.3889	0.6833	0.0167	0.6500	0.4500	0.3000	0.1667	**0.5333**	**0.5333**
		fluency	0.4000	0.3500	0.5167	0.4222	0.4500	0.4500	0.3000	0.4000	0.3500	0.3667	**0.5167**	0.3333
		relevance	0.7167	0.6000	0.6833	0.6667	0.6500	0.4333	0.7167	0.6000	0.4000	0.6000	0.4500	**0.8167**
	REALSumm	litepyramid	0.3913	0.5362	0.4420	0.4565	0.3551	0.3841	0.4420	0.3937	0.3261	0.3696	**0.5435**	**0.5435**
	NeR18	coherence	1.0000	1.0000	0.4286	0.8095	0.9048	0.9048	0.9048	0.9048	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		fluency	0.5238	1.0000	1.0000	0.8413	1.0000	0.5238	0.9048	0.8095	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		informativeness	1.0000	1.0000	1.0000	**1.0000**	0.7143	1.0000	0.9048	0.8730	0.7143	**1.0000**	**1.0000**	**1.0000**
		relevance	1.0000	0.5238	1.0000	0.8413	0.9048	0.9048	0.9048	0.9048	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	DialSummEval	consistency	0.7436	0.7179	0.4872	0.6496	0.7436	0.6410	0.6154	0.6667	0.5897	0.5641	0.5385	**0.7692**
		relevance	0.6923	0.4615	0.6410	0.5983	0.6410	0.6923	0.5385	0.6239	0.2308	0.4359	0.5897	**0.7179**
		fluency	0.5897	0.5641	0.5897	0.5812	0.3846	0.5641	0.5128	0.4872	0.1538	0.4872	**0.6410**	0.6154
		coherence	0.6667	0.7949	0.7436	0.7350	0.7436	0.7949	0.5897	0.7094	0.5897	0.6667	0.8205	**0.8974**
	OpenAI-axis1	accuracy	0.8000	0.0000	1.0000	0.6000	0.8000	1.0000	0.8000	0.8667	0.8000	0.0000	0.0000	**1.0000**
		coherence	0.4000	0.8000	0.0000	0.4000	0.8000	0.2000	0.8000	0.6000	**0.8000**	0.4000	0.2000	**0.8000**
		coverage	1.0000	1.0000	1.0000	**1.0000**	0.8000	0.8000	0.8000	0.8000	0.8000	**1.0000**	0.8000	0.8000
		overall	0.8000	1.0000	1.0000	0.9333	0.8000	1.0000	0.8000	0.8667	0.8000	**1.0000**	0.8000	**1.0000**
	OpenAI-axis2	accuracy	0.7143	0.4286	1.0000	0.7143	0.6190	0.7143	0.8095	0.7143	**1.0000**	0.5238	0.1429	0.9048
		coherence	0.2381	0.5238	0.3333	0.3651	-0.1429	0.2381	0.4286	0.1746	0.2381	**0.5238**	0.2381	0.4286
		coverage	1.0000	0.7143	0.9048	0.8730	1.0000	0.9048	1.0000	0.9683	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		overall	0.9048	0.7143	1.0000	0.8730	0.6190	1.0000	0.9048	0.8413	0.9048	0.9048	0.9048	**1.0000**
	OpenAI-CNN/DM1	accuracy	0.7333	0.8222	0.8222	0.7926	0.8667	0.7778	0.8222	0.8222	0.7333	0.6889	0.7778	**0.8667**
		coherence	0.5111	0.3333	0.5556	0.4667	0.4222	0.5111	0.5556	0.4963	0.5556	0.2000	**0.6000**	**0.6000**
		coverage	0.3778	0.3778	0.8667	0.5407	0.5111	0.8667	0.5111	0.6296	**1.0000**	**1.0000**	0.4222	0.8667
		overall	0.8667	0.5111	1.0000	0.7926	1.0000	0.7333	0.5111	0.7481	**1.0000**	0.3778	0.4667	**1.0000**
	OpenAI-CNN/DM3	accuracy	1.0000	0.3333	1.0000	0.7778	1.0000	0.3333	0.3333	0.5556	0.3333	**1.0000**	0.3333	**1.0000**
		coherence	1.0000	1.0000	1.0000	**1.0000**	1.0000	1.0000	1.0000	**1.0000**	**1.0000**	**1.0000**	0.3333	**1.0000**
		coverage	0.3333	1.0000	1.0000	0.7778	0.3333	1.0000	1.0000	0.7778	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		overall	0.3333	1.0000	1.0000	0.7778	0.3333	1.0000	1.0000	0.7778	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Task	Dataset	HE Metric	Random 1	Random 2	Random 3	Random Mean	Heuristic 1	Heuristic 2	Heuristic 3	Heuristic Mean	8M	SM	OL	CASF (ours)
MT	newstest2020 en-de	MQM	0.1429	0.1429	0.1429	0.1429	0.3333	0.1429	-0.0476	0.1429	0.1429	**0.3333**	0.1429	0.1429
		pSQM	0.8095	0.9048	0.9048	0.8730	0.8095	0.9048	0.9048	0.8730	**1.0000**	0.9048	0.9048	**1.0000**
	newstest2020 cn-en	MQM	0.7857	0.9286	0.7143	0.8095	0.6429	0.8571	0.7143	0.7381	0.1429	**0.9286**	0.8571	**0.9286**
		pSQM	0.4286	0.3571	0.7857	0.5238	0.2857	0.8571	0.4286	0.5238	0.3571	**0.9286**	0.7857	0.7857
	newstest2021 cn-en	MQM	0.0000	-0.1282	-0.0513	-0.0598	-0.0513	-0.0256	-0.0513	-0.0427	**0.4615**	0.1282	0.0000	0.0256
DialoGen	Persona Chat	Understandable	0.3333	-1.0000	0.3333	-0.1111	-1.0000	0.3333	0.3333	-0.1111	**0.3333**	**0.3333**	**0.3333**	**0.3333**
		Natural	0.3333	-1.0000	1.0000	0.1111	1.0000	-1.0000	0.3333	0.1111	0.3333	0.3333	0.3333	**1.0000**
		Maintains Context	1.0000	1.0000	1.0000	**1.0000**	1.0000	1.0000	1.0000	**1.0000**	-1.0000	**1.0000**	**1.0000**	**1.0000**
		Interesting	1.0000	1.0000	1.0000	**1.0000**	1.0000	0.3333	1.0000	0.7778	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		Uses Knowledge	1.0000	1.0000	1.0000	**1.0000**	-1.0000	1.0000	1.0000	0.3333	**1.0000**	**1.0000**	**1.0000**	**1.0000**
		Overall Quality	1.0000	1.0000	1.0000	**1.0000**	1.0000	1.0000	1.0000	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
StoryGen	MANS-ROC	overall	1.0000	1.0000	1.0000	**1.0000**	1.0000	1.0000	1.0000	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	MANS-WP	overall	1.0000	0.8000	0.8000	0.8667	0.8000	1.0000	1.0000	0.9333	**1.0000**	**1.0000**	**1.0000**	**1.0000**
MMGen	THUMB-MSCOCO	overall	1.0000	0.8000	1.0000	0.9333	1.0000	1.0000	1.0000	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	VATEX-EVAL	consistency	0.6000	1.0000	0.6000	0.7333	0.6000	1.0000	1.0000	0.8667	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Overall Performance			0.6880	0.6147	0.7503	0.6843	0.6149	0.6729	0.6547	0.7180	0.6686	0.7178	0.6790	**0.8332**

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Results and Analysis

Top-Ranked System Accuracy

- Experiment results on 137 real NLG evaluation setups with 44 human evaluation metrics across 16 datasets and 5 NLG tasks demonstrate the proposed method receives 93.18% top-ranked system recognition accuracy.

Table 2: Top-ranked accuracy on 16 datasets across 5 NLG tasks. 'Overall' represents the average result on all human metrics from all tasks. **Bold number** indicates that the method has the best performance among all methods.

Method	SUM	MT	DialoGen	StoryGen	MMGen	Overall
Random	0.7586	0.8667	0.7778	0.6667	1.0000	0.7597
Heuristic	0.8046	0.6667	0.7778	0.6667	1.0000	0.7829
8M	0.8276	0.8000	0.8333	1.0000	1.0000	0.8409
SM	0.8966	1.0000	0.8333	1.0000	1.0000	0.9091
OL	0.6897	0.8000	1.0000	1.0000	1.0000	0.7727
CASF (ours)	0.9310	0.8000	1.0000	1.0000	1.0000	**0.9318**

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Results and Analysis

PEKING UNIVERSITY

Case Study

- The risk of random sampling: Different sampling subsets may result in different inter-system rankings, making human evaluation unreliable.
- CASF selects the same subset in multiple times of sampling, and the variance of the inter-ranking accuracy obtained by multiple sampling times on a total of 44 human metrics is 0.
- Since CASF selects representative samples, it obtains more accurate inter-system rankings, making human evaluation more reliable.

System Ranking / Sampling Method \ System	sup4_ppo_rm4_t.7	pretrain_6b_t.7	sup4_6b_ppo_rm4_6b_t.7	sup4_6b_t0.7	sup4_12b_t0	Kendall's Tau
Ground Truth	3	4	1	0	2	
Random 1	4	3	1	0	2	0.80
Random 2	3	1	4	0	2	0.00
Random 3	1	3	4	0	2	0.20
Heuristic 1	1	4	3	0	2	0.40
Heuristic 2	1	3	4	0	2	0.20
Heuristic 3	4	3	1	0	2	0.80
8M	4	3	1	0	2	0.80
SM	3	1	4	0	2	0.00
OL	3	1	4	0	2	0.00
CASF (ours)	3	4	1	0	2	**1.00**

Figure 4: Inter-system ranking of human evaluation aspect 'accuracy' of the summarization dataset OpenAI-axis1. Ground truth is the inter-system ranking on the entire dataset. Other sampling methods take 50% of the dataset. Rankings in red indicate incorrect rankings.

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Results and Analysis

Automatic Metric for Preliminary Phase

Table 4: Experiment results of CASF on NLG tasks pre-ranking on different automatic metrics. 'Overall' represents the average result on all human metrics from all tasks. **Bold number** indicates that the automatic metric ranks first among all automatic metrics. <u>Underlined number</u> indicates that the automatic metric ranks second among all metrics.

Automatic Metric	SUM	MT	DialoGen	StoryGen	MMGen	Overall
BERT-SCORE	0.7361	0.5799	0.6667	1.0000	1.0000	0.7329
MOVER-SCORE	0.8429	0.5766	0.8889	1.0000	1.0000	**0.8332**
ROUGE-1	0.7308	0.5700	0.6667	0.3000	1.0000	0.6965
ROUGE-2	0.7266	0.5491	0.5556	1.0000	0.8000	0.6989
ROUGE-L	0.7196	0.5209	0.8889	1.0000	1.0000	0.7456
BART-SCORE	0.6015	0.4370	0.8889	0.9000	0.8000	0.6446
BLEU	0.7196	0.3718	0.5556	1.0000	0.8000	0.6741
METEOR	0.7821	0.5359	0.8889	1.0000	1.0000	<u>0.7885</u>

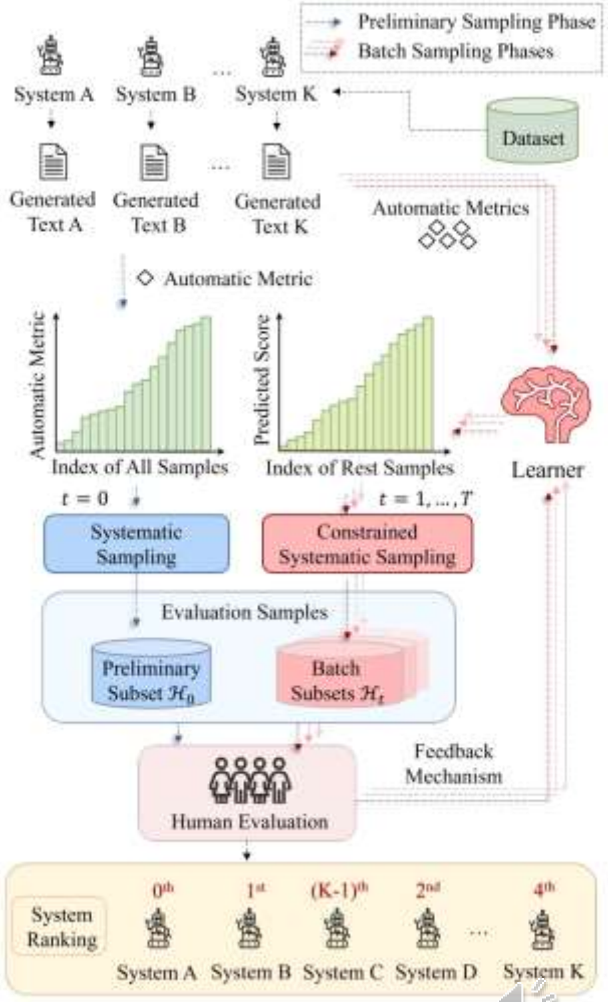


Fig. Constrained Active Sampling Framework

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Results and Analysis

Phases and Associated Sampling Ratios

- In most cases, the experimental performance is better when the number of iteration phases is 5.
- There is no need to set the preliminary sampling ratio and the batch sampling ratio separately, because it is simple and effective to directly sample each phase according to the total sampling rate and the number of phases.

Table 3: Experimental results on 44 human metrics with different mode (Average and Preliminary-Fixed (P-Fixed)), number of phases (# Phase), preliminary sample ratio (P-R) and batch sampling ratio (B-R) of each phase for the proposed CASF.

Mode	# Phase	P-R	B-R	Tau	Mode	# Phase	P-R	B-R	Tau	Mode	# Phase	P-R	B-R	Tau	Mode	# Phase	P-R	B-R	Tau
	2	0.2500	0.2500	0.7507		2	0.1000	0.4000	0.7330		2	0.0500	0.4500	0.7449		2	0.1500	0.3500	0.7347
	3	0.1667	0.1667	0.7567		3	0.1000	0.2000	0.7547		3	0.0500	0.2250	0.7418		3	0.1500	0.1750	0.7670
	4	0.1250	0.1250	0.7557		4	0.1000	0.1333	0.8046		4	0.0500	0.1500	0.7663		4	0.1500	0.1167	0.7612
	5	0.1000	0.1000	**0.8332**		**5**	0.1000	0.1000	**0.8332**		**5**	0.0500	0.1125	**0.7739**		5	0.1500	0.0875	0.7647
Average	6	0.0833	0.0833	0.7214	P-Fixed	6	0.1000	0.0800	0.7543	P-Fixed	6	0.0500	0.0900	0.7276	P-Fixed	6	0.1500	0.0700	0.7109
	7	0.0714	0.0714	0.7237		7	0.1000	0.0667	0.6892		7	0.0500	0.0750	0.6859		7	0.1500	0.0583	0.7285
	8	0.0625	0.0625	0.7037		8	0.1000	0.0571	0.7269		8	0.0500	0.0643	0.7158		**8**	0.1500	0.0500	**0.7884**
	9	0.0556	0.0556	0.7258		9	0.1000	0.0500	0.7237		9	0.0500	0.0563	0.7190		9	0.1500	0.0438	0.7471
	10	0.0500	0.0500	0.7511		10	0.1000	0.0444	0.7250		10	0.0500	0.0500	0.7511		10	0.1500	0.0389	0.6987

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

PART 05

Conclusion

Conclusion

北京大學 PEKING UNIVERSITY

Towards Reliable Human Evaluation

- We focused on giving a more correct inter-system ranking for reliable human evaluation with limited time and cost.
- We propose a Constrained Active Sampling Framework and show the overall inter-system Kendall correlation improved by 41% to 0.83 compared to the widely used random sampling method in a total of 44 human evaluation metrics across 16 datasets in 5 NLG tasks. CASF ranked first or ranked second among all comparison methods on up to 90.91% of the human metrics.
- We release a tool and we strongly recommend using the Constrained Active Sampling Framework for reliable human evaluation in future works to get a more reliable inter-system ranking.

[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

Acknowledgement

PEKING UNIVERSITY

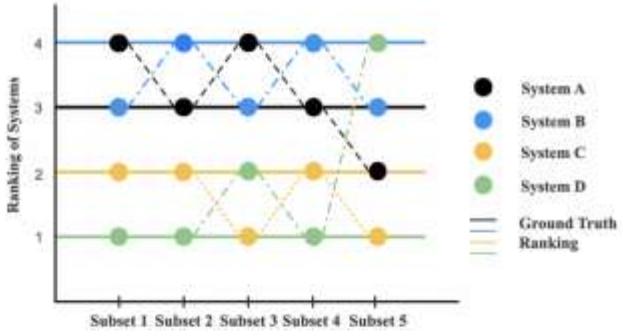
Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, Yuesheng Zhu

Peking University
{ruanjie,puxiao}@stu.pku.edu.cn, {gaomingqi,wanxiaojun,zhuys}@pku.edu.cn

Abstract

Human evaluation is viewed as a reliable evaluation method for NLG which is expensive and time-consuming. In order to save labor and costs, researchers usually perform human evaluation on a small subset of data sampled from the whole dataset in practice. However, different selection subsets will lead to different rankings of the systems. To give a more correct inter-system ranking and make the gold standard human evaluation more reliable, we propose a Constrained Active Sampling Framework (CASF) for reliable human judgment.



[1] **Jie Ruan**, Xiao Pu, Minqi Gao, Xiaojun Wan and Yuesheng Zhu. Better than Random: Reliable NLG Human Evaluation with Constrained Active Sampling. Accepted by AAAI 2024.

北京大學
PEKING UNIVERSITY

THANK YOU!

Better than Random:
Reliable NLG Human Evaluation with Constrained Active Sampling

Presenter: Jie Ruan

ruanjie@stu.pku.edu.cn