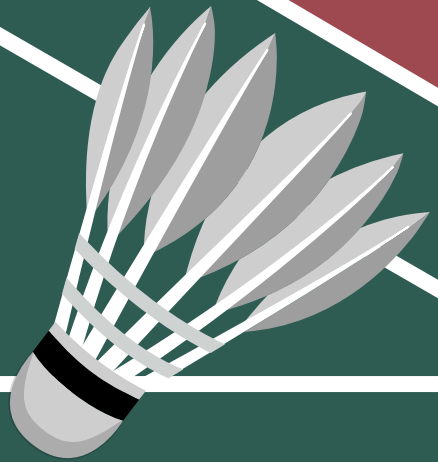
2024/08/04
IT4PSS, IJCAI 2024

BADGE: BADminton report Generation and Evaluation with LLM



Shang-Hsuan Chiang, Lin-Wei Chao, Kuang-Da Wang, Chih-Chuan Wang, Wen-Chih Peng

TABLE OF CONTENTS

01 Introduction

Introduction

Motivation

Badminton reports generally include details such as player names, game scores, and ball types, providing audiences with a comprehensive view of the games.

Challenge

Manually writing these reports can be subjective and time-consuming.

Objective

Explore whether a Large Language Model (LLM) could automate the generation and evaluation of badminton reports.

Badminton Report

Kento Momota was the top seed but he took over the world #1 spot from Viktor Axelsen in late September. He came in with a 3-match winning streak over Chinese Taipei's **Chou Tien Chen** but the world #4 won their first meeting in 2018 en route to winning the German Open.

Chou's stamina issues were not helped by the end of the first game. After the two players battled to **20-all**, Kento Momota hit an impossibly good **net tumble** then was not faulted even though the reply clearly showed he'd reached over the net to **kill** a reply from Chou that was not going to even make it over the net. Momota beat Chou on the same **front forehand corner** one rally later and claimed the first game **22-20**.

…

Research Questions

1. What is the best configuration for generating badminton reports?
2. How to automatically evaluate the quality of badminton reports?
3. What is the difference between LLM-generated and human-written badminton reports?

02 Related Works

Badminton Dataset

ShuttleSet [Wang et al., 2023b]

- Shot trajectories
- Rally duration
- Point outcomes
- Player names
- Tournament settings

Generation with LLM

In-Context Learning

- Zero-shot
- One-shot
- Few-shot [Brown et al., 2020]
- Chain of Thought (CoT) [Wei et al., 2022]

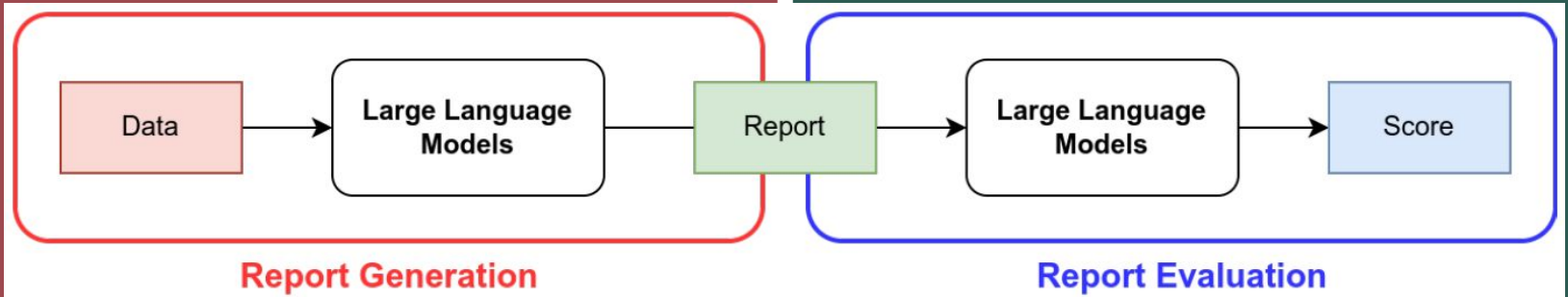
Evaluation with LLM

G-Eval [Liu et al., 2023]

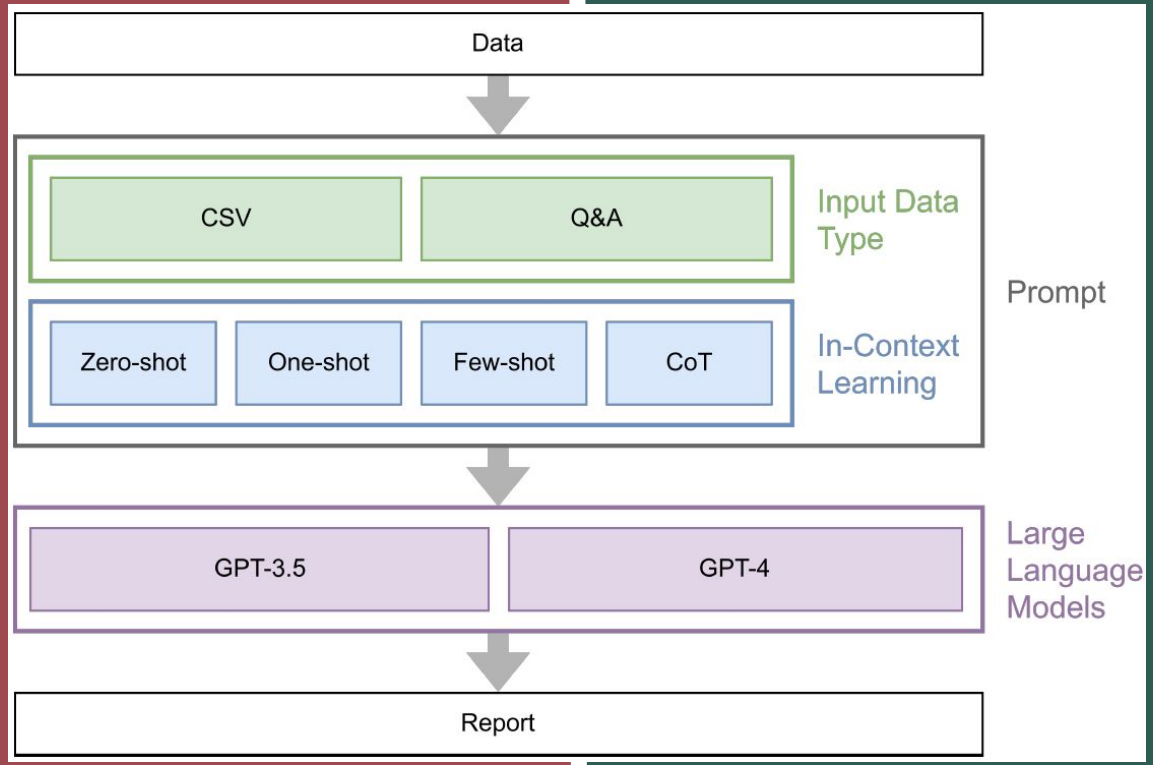
Encompasses chain-ofthought and weighting techniques for assessing the coherence, consistency, and fluency of news summaries.

03 Methods

Overview


Data
Large Language Models
Report
Large Language Models
Score
Report Generation
Report Evaluation

Report Generation



Input Data Type

CSV:
win_point_player, win_reason, ball_types, lose_reason, roundscore_A, roundscore_B
Ratchanok Intanon, opponent goes out of bounds, lob, goes out of bounds, 0, 1
An Se Young, opponent hits the net, push, hits the net, 1, 1
Ratchanok Intanon, wins by landing, smash, opponent wins by landing, 1, 2
...

CSV
(structured and rally-level data)

Q&A:
Q1: Which player won the game? How many points did the winner get?
A1: An Se Young won the game with 22 points.
Q2: Which player lost the game? How many points did the loser get?
A2: Ratchanok Intanon lost the game with 20 points.
...

Q&A
(unstructured and set-level data)

In-Context Learning (ICL)

Zero-shot:
You are a reporter for badminton games.
...

Few-shot:
You are a reporter for badminton games.
...
I give you some example reports as reference:
Example 1:
...
Example 2:
...

One-shot:
You are a reporter for badminton games.
...
I give you an example report as a reference:
Example:
...

CoT:
You are a reporter for badminton games.
...
Let's think step by step:
1. Read the CSV table carefully and understand this badminton game.
2. ...

Large Language Models (LLM)

GPT-3.5 [OpenAI, 2022]
(GPT-3.5-turbo-0125)

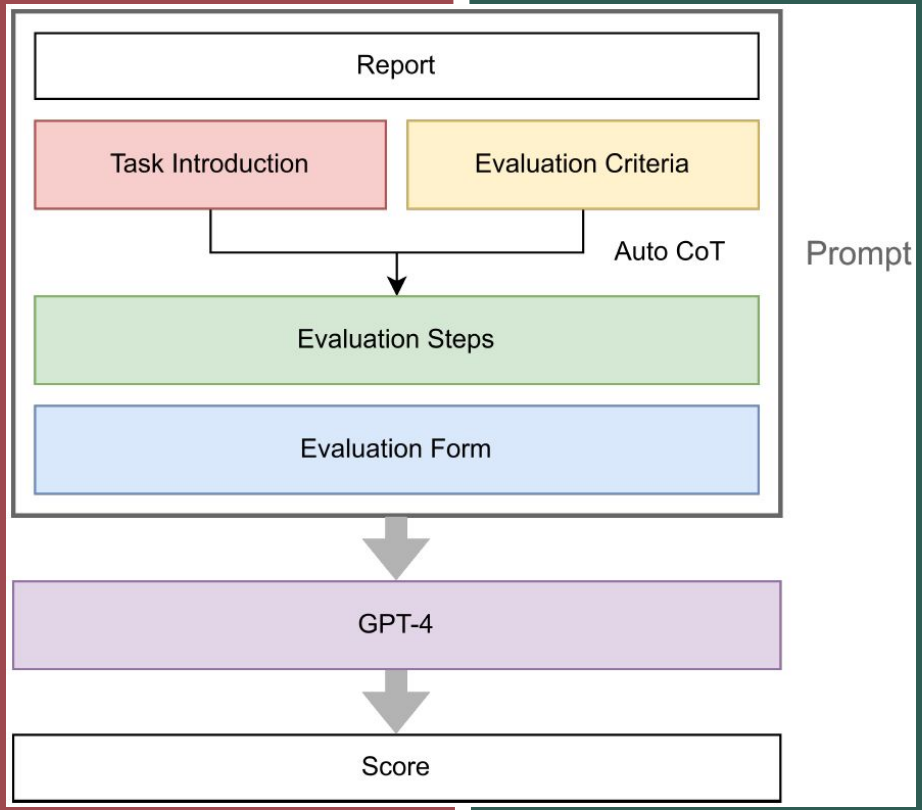
GPT-4 [Achiam et al., 2023]
(GPT-4-turbo-2024-04-09)

Report Evaluation

GPT-4 Evaluation

Human Evaluation

GPT-4 Evaluation



G-EVAL
[Liu et al., 2023]

GPT-4 Evaluation

Task Introduction:
You are a reviewer of the badminton reports.
I will give a badminton report, please follow the Evaluation Steps to score this badminton report based on the Evaluation Criteria.
...

- **Coherence (1-10):** means being logical and clear in thought or communication, where ideas fit together smoothly to form a unified whole.
- **Consistency (1-10):** refers to the quality of being steadfast, reliable, and uniform in behavior, performance, or appearance over time.
- **Excitement (1-10):** is a feeling of enthusiasm or thrill, often before or during an event or activity.
- **Fluency (1-10):** the quality of the summary in terms of grammar, spelling, punctuation, word choice, and sentence structure.

Evaluation Steps:
1. Read for Structure and Organization: ...
2. Sentence-Level Analysis: ...
3. Overall Coherence Assessment: ...

Auto CoT

Human Evaluation

For the human evaluation, we prepared a form containing three badminton reports authored by GPT-3.5, GPT-4, and humans, respectively.

04 Experiments

Dataset

ShuttleSet
[Wang et al., 2023b]

We sample 10 badminton games spanning the years 2018 to 2021

5 men's singles
5 women's singles

We only extract the 6 most crucial columns, which include win point player, win reason, lose reason, ball types, roundscore A, and roundscore B.

Result for Input Data Type

CSV > Q&A

Data Type + ICL	Coherence	Consistency	Excitement	Fluency	Avg.
CSV + zero-shot	8.2	*7.5*	7.9	8.8	*8.100*
CSV + one-shot	8.4	8.3	7.8	8.8	8.325
CSV + few-shot	8.3	9.0	7.7	8.7	8.425
CSV + CoT	8.4	**9.2**	**8.0**	**8.9**	**8.625**
Q&A + zero-shot	*7.9*	8.6	*7.3*	8.7	8.125
Q&A + one-shot	**8.6**	8.4	7.4	8.8	8.300
Q&A + few-shot	8.3	8.5	7.5	8.6	8.225
Q&A + CoT	*7.9*	8.7	7.4	*8.5*	8.125

Result for Input Data Type

CSV data type are more prone to hallucinations!

Ground Truth	CSV	Q&A
21-19	In the deciding Set 3, Gemke fought back with determination, forcing Ginting to work hard for every point. Despite Gemke's efforts, Ginting managed to maintain his composure and clinch the set 21-21, securing his spot in the semifinals.	However, Gemke made a comeback in the third set, defeating Ginting with a score of 21-19. Gemke's consistent play and ability to keep the shuttle in play proved to be crucial in securing the set.

Result for In-Context Learning (ICL)

CoT > Few-shot > One-shot > Zero-shot

Data Type + ICL	Coherence	Consistency	Excitement	Fluency	Avg.
CSV + zero-shot	8.2	*7.5*	7.9	8.8	*8.100*
CSV + one-shot	8.4	8.3	7.8	8.8	8.325
CSV + few-shot	8.3	9.0	7.7	8.7	8.425
CSV + CoT	8.4	**9.2**	**8.0**	**8.9**	**8.625**
Q&A + zero-shot	*7.9*	8.6	*7.3*	8.7	8.125
Q&A + one-shot	**8.6**	8.4	7.4	8.8	8.300
Q&A + few-shot	8.3	8.5	7.5	8.6	8.225
Q&A + CoT	*7.9*	8.7	7.4	*8.5*	8.125

Result for Large Language Models (LLM)

GPT-4 > GPT-3.5 > Humans

Writer	Coherence	Consistency	Excitement	Fluency	Avg.
Human	*7.5*	*8.9*	*6.8*	*8.5*	*7.925*
GPT-3.5	8.4	9.2	8.0	8.9	8.625
GPT-4	**8.6**	**9.4**	**8.2**	**9.1**	**8.825**

Result for Human Evaluation

GPT-4 > Humans > GPT-3.5

Writer	Coherence	Consistency	Excitement	Fluency	Avg.
Human	7.6	7.5	6.9	7.8	7.450
GPT-3.5	*6.5*	*7.3*	*5.2*	*6.4*	*6.350*
GPT-4	**8.3**	**8.2**	**8.0**	**8.4**	**8.225**

05 Limitations & Future Works

Limitations

Badminton report generation is a relatively unexplored topic in the research field, leaving us without other baselines for comparison.

Lack a quantitative method to measure the occurrence of hallucinations in the reports.

The bias that GPT-4 prefers the reports generated by LLM may lead to unfair evaluation.

Future Works

Constructing a benchmark (comprising dataset and evaluation metrics).

Employing a Q&A model to extract answers from reports.

Exploring solutions to this issue represents a promising direction for future research.

06 Conclusion

Takeaways

BADGE

A pioneering framework for badminton report generation and evaluation.

Generation

We found that reports generated by GPT-4 with CSV and Chain of Thought exhibit the best performance.

Evaluation

Revealing a bias where GPT-4 prefers reports generated by LLMs.

SCAN ME

Paper

THANK YOU!

DO YOU HAVE ANY QUESTIONS?

Shang-Hsuan Chiang
Department of Computer Science,
National Yang Ming Chiao Tung University, Hsinchu, Taiwan
andy10801@gmail.com

SCAN ME

Demo Website

Thanks For Listening!