



# Trends in LLM Collaboration

Debaters and Judges in  
Agentic AI Systems

# Using LLMs (Zero-shot)



Human

Please write an article about {topic}

LLM response ...



LLM

# Using LLMs (Zero-shot)

- Write an email
- Draft an article
- Extract information from an essay
- Etc.



LLM



# More Advanced LLM usage



Human

First read my new emails and then send a summary as a slack message, first provide the plan and then take action using the available tools

1. Tool: Gmail  
read new emails
2. Summarize
3. Tool: Slack  
send summary to channel #new-emails



LLM

Call  
Response

Tools



# Use Agents (Simple Example)



Human

Build a Python app that achieves the following goals:

- Find restaurants in the Halifax and Dartmouth areas.
- Retrieve their menus and store them in a database.
- Use a search engine to find the relative calorie and protein information for each menu item and store the data in the database.



LLM

Action



Feedback



Tools:

- Search / retrieve  
get relevant  
information



- Python Interpreter  
Execute code



- Database  
Store information





**Software Developer**



coding



testing



Deployment

**Overwhelmed and  
less efficient**



**Software Developer**  
coding



**Test Engineer**  
testing



**DevOps**  
Deployment

**Each specialist focuses on  
their expertise**

**Faster, higher-quality results**



# Use Multiple Agents (Example: Chatdev)



Agents take on roles such as

- CEO
- CTO
- Programmer
- Tester
- Reviewer
- Designer

# Recent Trends?

- **Collaborate**  
Two or more LLMs collaborate to solve a problem
- **Debate**  
Few LLMs debate until achieve a conclusion
- **LLM-as-a-Judge**  
An LLM Judges it's own generation or others



# Collaborate



Human

You will be given a string of words separated by commas or spaces. Your task is to split the string into words and return an array of the words.

```
def task():  
.....
```



LLM 1

```
def task():  
....
```

There is a bug  
on line 4

Thank you for the  
feedback, here is the  
revised code:  
def task(): ....



LLM 2

## Collaborate (ensemble)



Human

Who discovered the law of gravity?



LLM 1

Isaac Newton



LLM 2

Newton



LLM 3

Isaac Newton

# Debate



Human

What are the best practices for sustainable living?

LLM 1



Reducing waste ...

LLM 2



Conserving energy ....

LLM 3



Using eco friendly products...



...

# LLM-as-a-judge



Human

You will be given a string of words separated by commas or spaces. Your task is to split the string into words and return an array of the words.

```
def task():  
    ....
```



LLM

Here are two approaches to solving ...

Approach one is better because ...



LLM-as-a-Judge

# LLM-as-a-Judge beginning



LLM-as-a-Judge

---

## Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

---

Lianmin Zheng<sup>1\*</sup> Wei-Lin Chiang<sup>1\*</sup> Ying Sheng<sup>4\*</sup> Siyuan Zhuang<sup>1</sup>

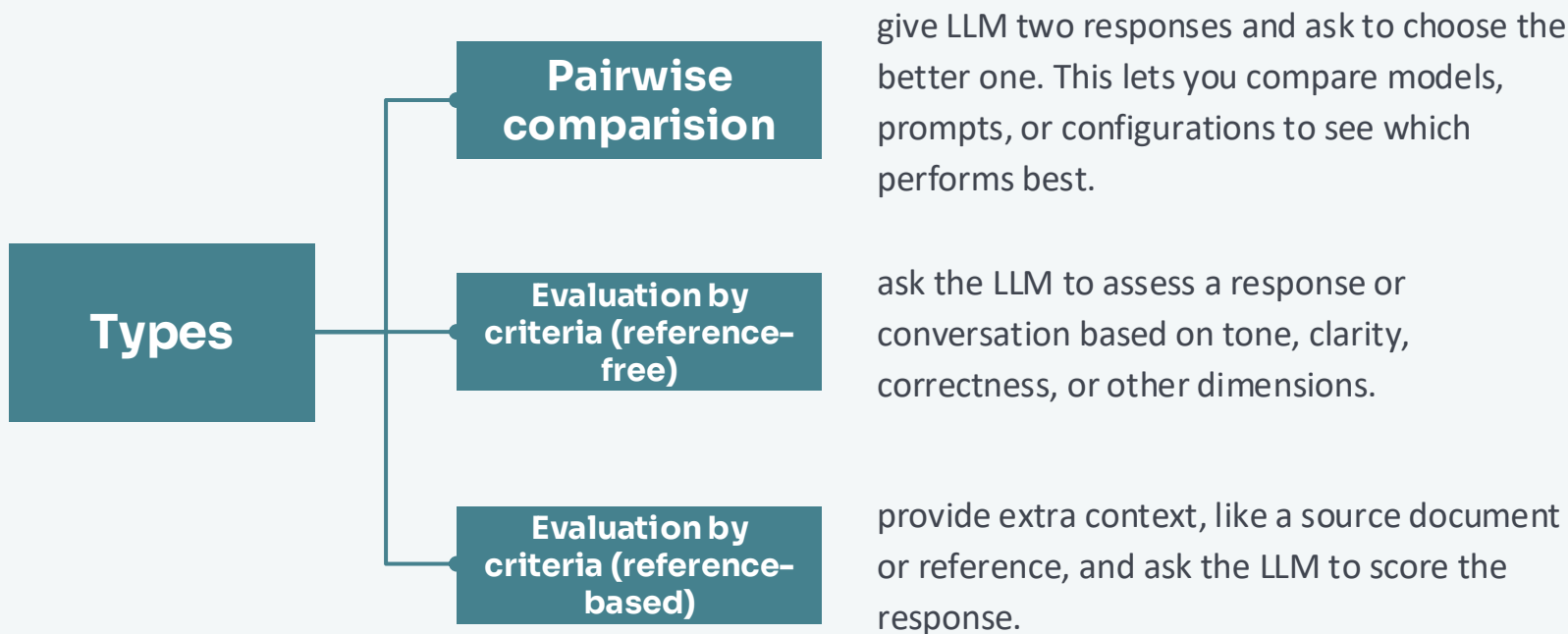
Zhanghao Wu<sup>1</sup> Yonghao Zhuang<sup>3</sup> Zi Lin<sup>2</sup> Zhuohan Li<sup>1</sup> Dacheng Li<sup>13</sup>

Eric P. Xing<sup>35</sup> Hao Zhang<sup>12</sup> Joseph E. Gonzalez<sup>1</sup> Ion Stoica<sup>1</sup>

<sup>1</sup> UC Berkeley   <sup>2</sup> UC San Diego   <sup>3</sup> Carnegie Mellon University   <sup>4</sup> Stanford   <sup>5</sup> MBZUAI

Dec 2023

# Types of LLM-as-a-judge



# Pairwise Comparison

LLM 1



Response A

LLM 2



Response B



LLM-as-a-Judge

Response B is better,  
because ...

Which response is better?  
Criteria: ....

## Evaluation by criteria (reference free)

LLM 1



Response A

LLM 2



Response B

LLM-as-a-Judge



Response  
A is better,  
because ...

**Evaluate which response  
has a good tone?**

Criteria: ....

**Evaluate which response  
is more polite?**

Criteria: ....



## Reference based: Evaluating correctness based on reference answer



LLM

Respond A

Yes.



LLM-as-a-Judge

Is the response correct  
compared to the  
reference?  
Criteria: ....

Reference response: ....

# Reference based: Evaluating answer quality considering the question



LLM

What are the best practices for sustainable living?

Conserving energy ....



Human

Does this respond fully answer the question?  
Criteria ....

No.

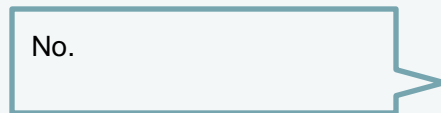
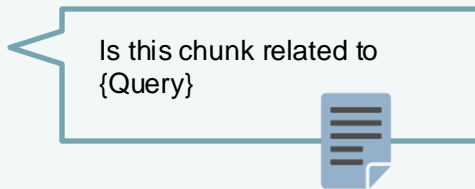


LLM-as-a-Judge

# Reference based: Scoring context relevance in RAG



Human



All context chunks



LLM-as-a-Judge

# Reference based: Evaluating hallucinations in RAG



Human

Is the {Response} relevant  
to the context?  
Criteria ...



Yes.



All context chunks



LLM-as-a-Judge

# Practical Strategies for Implementing LLMs: From Concept to Application.

## 1. Start Simple

**Don't Use Collaboration, Debate, LLM-as-a-Judge or Agent!**  
**If your problem is not complex!**

**2. Different models like to be prompted differently, try many variations before giving up and implementing complex approaches!**

# **Practical Strategies for Implementing LLMs: From Concept to Application.**

## **3. Don't Reinvent the Wheel!**

**There are many research out there about variety of use cases make sure to not the same mistakes and learn effective strategies!**

## **4. Always evaluate!**

**Eyeballing does not mean your system works! Get some ideas from LLM-as-a-Judge framework and start simple!**

# Thank you! Let's Connect!

LinkedIn:



Slides:

<https://www.linkedin.com/in/fatemehrahimi/>

# Demo

Short Description of the demo

