Trends in LLM Collaboration

Debaters and Judges in Agentic AI Systems

Fateme Rahimi – January 2025

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
- Debug a code
- Etc.

Qwen

MISTRAL AI_

LLaMA by Meta

LLM

ChatGPT

Gemini

Claude

Aya

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

Gmail

slack

Recommend Reading: Building effective agents by Anthropic

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

tavily

- Python Interpreter
  Execute code

python

- Database
  Store information

mongoDB.

Recommend Reading: Building effective agents by Anthropic

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

Image Source: ChatDev: Communicative Agents for Software Development, Qian et al.

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 – Example: pair-programming

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

Recommend Reading: Merge, Ensemble, and Cooperate! A Survey on Collaborative Strategies in the Era of Large Language Models, Lu et al. 2024

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 echo friendly products...

...

Recommend Reading: Multi-LLM Debate: Framework, Principals, and Interventions, Estornell et al. 2024

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

Question Source: HumanEval

LLM-as-a-Judge beginning

LLM-as-a-Judge

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

Lianmin Zheng[1*] Wei-Lin Chiang[1*] Ying Sheng[4*] Siyuan Zhuang[1]

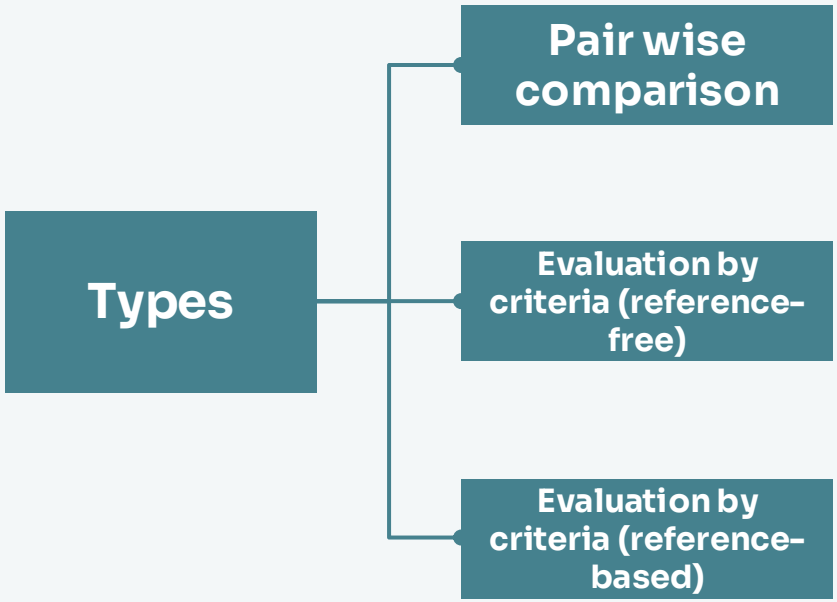
Zhanghao Wu[1] Yonghao Zhuang[3] Zi Lin[2] Zhuohan Li[1] Dacheng Li[13]

Eric P. Xing[35] Hao Zhang[12] Joseph E. Gonzalez[1] Ion Stoica[1]

[1] UC Berkeley [2] UC San Diego [3] Carnegie Mellon University [4] Stanford [5] MBZUAI

Dec 2023

Types of LLM-as-a-judge



give LLM two responses and ask to choose the better one. This lets you compare models, prompts, or configurations to see which performs best.

ask the LLM to assess a response or conversation based on tone, clarity, correctness, or other dimensions.

provide extra context, like a source document or reference, and ask the LLM to score the response.

Pairwise Comparison

Which response is better?

LLM 1

Response A

Response B is better, because ...

LLM-as-a-Judge

LLM 2

Response B

Evaluation by criteria (reference free)

Evaluate which response has a good tone?
Criteria: ….

Evaluate which response Is more polite?
Criteria: ….

LLM 1

Response A

LLM-as-a-Judge

Response A is better, because ...

LLM 2

Response B

Reference based: Evaluating correctness based on reference answer

LLM

Response.

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

Is this chunk related to {Query}

No.

LLM-as-a-Judge

Reference based: Evaluating hallucinations in RAG

Human

All context chunks

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

Yes.

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. Build strategically!

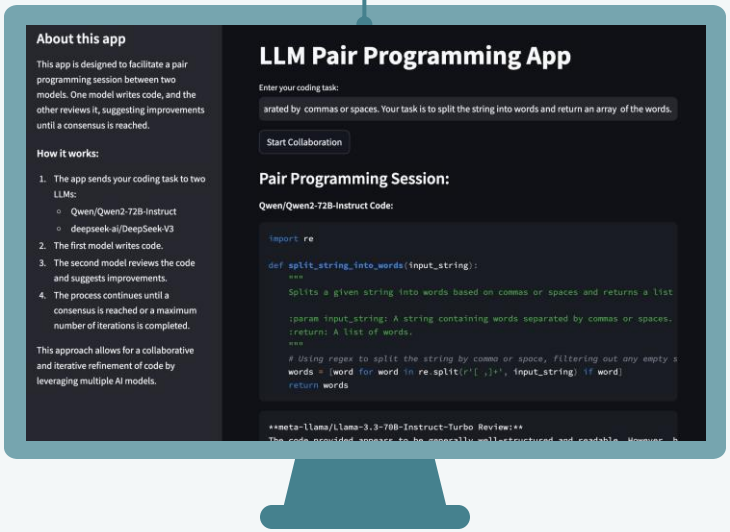
There are many research out there about variety of use cases make sure to not the same mistakes and learn effective strategies!

4. Always evaluate!

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

Demo

Let's look at an example of collaboration and LLM-as-a-judge



Thank you! Let's Connect!

LinkedIn:

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

Slides:

https://fatemerhmi.github.io/coffee-gen-ai/talks/Trends_in_LLM_Collaboration-Fatemeh_Rahimi-Jan_2025.pdf