



# Automatic Evaluation of Conversational AI Chatbots Using Large Language Models



Full paper link:  
<https://tinyurl.com/chatbot-llm>



Xiaoyi Tian\*

NC STATE  
UNIVERSITY



Yukyeong Song

UF UNIVERSITY of  
FLORIDA



Amogh Mannekote

UF UNIVERSITY of  
FLORIDA



Maya Israel

UF UNIVERSITY of  
FLORIDA



Kristy Elizabeth Boyer

UF UNIVERSITY of  
FLORIDA

# Introduction

- Project-based learning is increasingly utilized in STEM and artificial intelligence (AI) education
  - increased engagement (Kokotsaki, 2016)
  - deeper understanding of complex concepts (Guo, 2010)
- Major challenge of project-based learning
  - evaluating learner projects
  - providing timely feedback
  - **time-consuming and resource-intensive**
- LLM might help
  - They have shown promising results in grading short answers (Funayama, 2023) and evaluating essays (Mizumoto, 2023)



What about using LLMs to assess computational artifacts (both technical and creative aspects)?

Chatbot  
Development  
Platform:  
**AMBY**

The screenshot displays the AMBY chatbot development interface, divided into three main sections: **Top-level Intents**, **Follow-up Intents**, and **Testing Window**.

**Top-level Intents:** A central 'USER' icon branches into four main categories: 'North America', 'Help', 'Europe', and 'Greet'. Below these are 'Default Fallback' and '+ ADD AN INTENT' buttons.

**Follow-up Intents:** This section shows specific follow-up intents for each top-level category. 'North America' includes 'USA', 'Canada', and 'Mexico'. 'Europe' includes 'Italy', 'France', and 'Spain'. 'Greet' is linked to a 'Greet Intent' box, and 'Default Fallback' is linked to a 'Default Fallback Intent' box.

**Testing Window:** This section shows a simulated chat interaction. The user asks, 'I want to learn about africa'. The chatbot responds with a detailed paragraph about Africa. The user then asks, 'tell me more about nigeria', and the chatbot provides another detailed paragraph about Nigeria. The interface includes a 'CHAT' button, a user profile for 'MRWORLDWIDE', and a 'Type Something!' input field.

# Training Phrases

STACKED SIDEBYSIDE Impact on Oceans ✕

< Training Phrases

Example sentences for the agent to understand the user's intent. At least 3 training phrases required.

ADD ▶

Can you explain the impact of climate change on the oceans

Does climate impact the oceans?

How does it impact the sea?

What are some potential impacts on Oceans?

TRAIN THE AI ▶

# Responses

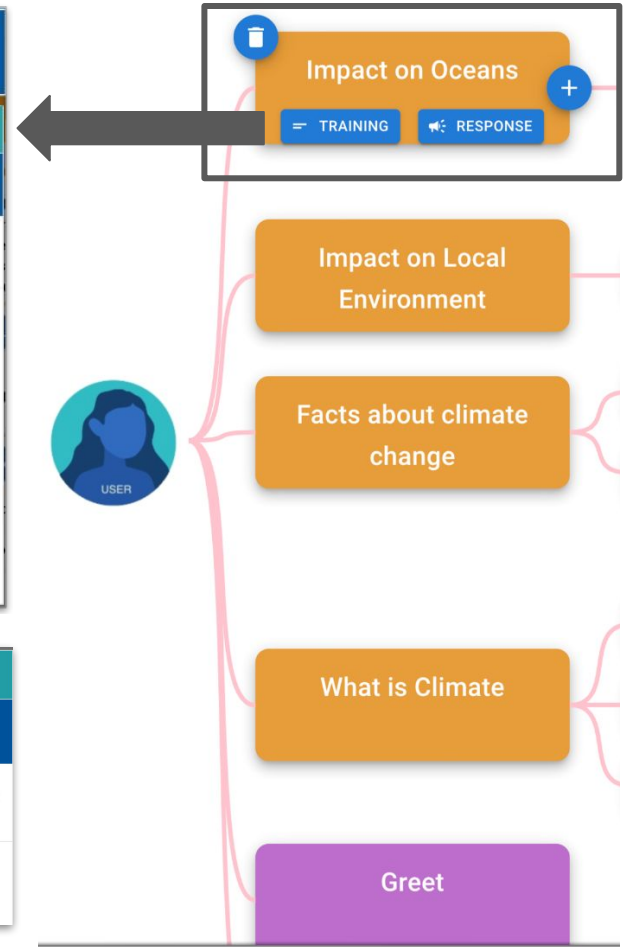
< Responses

A list of response that the agent will select from the intent, Impact on Oceans. At least 1 response required.

ADD ▶

There are many impact on oceans, including melted ice, increased sea level and ocean acidification.

# Intents





# Research Questions

- RQ1: How do LLMs perform in assessing different aspects of computational artifacts?
- RQ2: What are the tradeoffs among different prompting strategies?




# Context: Middle School AI Summer Camp

- Two-week middle school AI summer camps over two years (Katuka et al., 2023; Song et al., 2023)
  - general CS and AI lessons
  - conversational AI (AMBY) lessons
  - unplugged activities
  - chatbot project development
- 75 chatbot projects collected
  - 66 created by middle school learners, 9 by undergraduate learners during pre-camp workshop



# Chatbot Artifact Rubric Dimensions

1. Greet intent
2. Default fallback intent
3. Follow-up intents
4. Training phrases
5. Responses



Each dimension rated as 1-4  
Rubric Cohen's kappa = 0.82



# Chatbot Artifact Rubric Dimensions

1. Greet intent
2. Default fallback intent
3. Follow-up intents
4. Training phrases
5. Responses

Conversational  
Design

AI  
Development

Each dimension rated as 1-4  
Rubric Cohen's kappa = 0.82

# Chatbot Artifact Rubric Statements

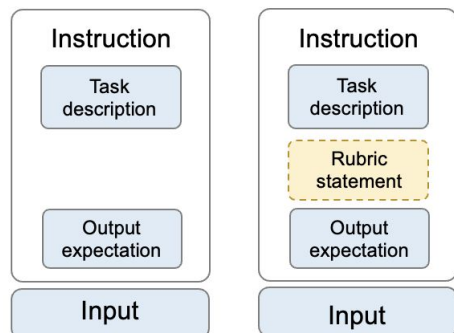
<b>Artifact Dimensions</b>	<b>Statement for Score of 3 (Meeting expectations)</b>
<b>Greet intent</b>	At least one customized greet response demonstrating its purpose. May not set exact user expectations.
<b>Default fallback intent</b>	At least one customized default fallback response that can redirect the users.
<b>Follow-up intents</b>	Multiple logical follow-up intents. Each follow-up intent is related to its parent intent mostly logically and can be triggered properly based on the responses from their parent intents.
<b>Training phrases</b>	Most training phrases are ample, cohesive, and varied within the intent.
<b>Responses</b>	At least one response is of appropriate length, logical, conversational, and mostly free from grammatical errors.

# LLM-based Project Assessment Implementation

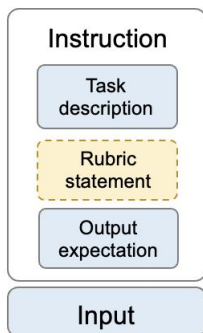
- LLM<sub>4</sub>Qual open-source framework for experiment
  - [github.com/msamogh/llm4qual](https://github.com/msamogh/llm4qual)
- GPT<sub>4</sub> (state of art LLM *in Jan 2024*)
- Four prompting strategies:
  - zero-shot-basic
  - zero-shot-rubric
  - few-shot-basic
  - Few-shot-rubric
- Data Splits: training, validation, testing
- Prompt Engineering

# Prompt Strategies

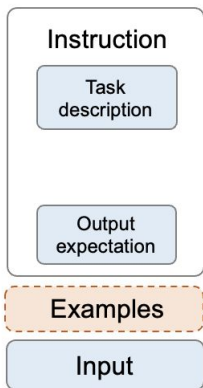
# Prompt Template



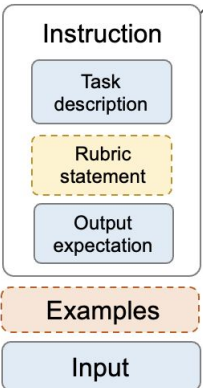
**Zero-shot-basic**



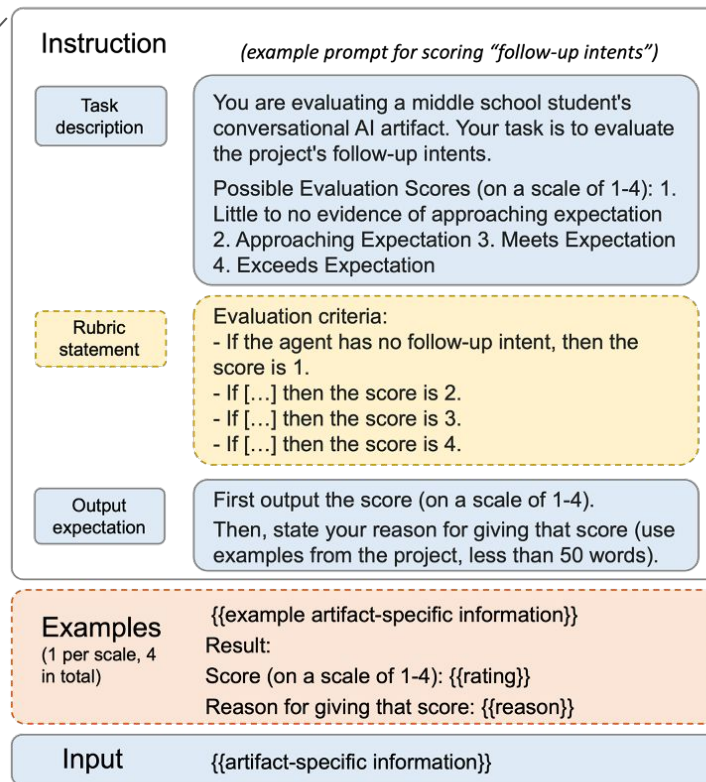
**Zero-shot-rubric**



**Few-shot-basic**



**Few-shot-rubric**



**Few-shot-rubric**

# Evaluation Metrics

- Human-GPT<sub>4</sub> alignment
  - Spearman correlation ( $\rho$ )
  - Weighted Cohen's Kappa (QWK)

# Results

Artifact Dimensions	Metrics	Human-human	Human-GPT4			
			Zero-shot	Zero-shot	Few-shot	Few-shot
			Basic	Rubric	Basic	Rubric
Greet intent	$\rho$	0.850	0.339	0.641	<b>0.659</b>	0.646
	QWK	0.820	0.325	0.623	<b>0.698</b>	0.645
Default Fallback intent	$\rho$	0.979	0.179	0.782	0.779	<b>0.816</b>
	QWK	0.984	0.252	0.750	0.781	<b>0.797</b>
Follow-up intents	$\rho$	0.839	0.133	0.217	0.203	<b>0.346</b>
	QWK	0.805	0.154	0.244	0.230	<b>0.388</b>
Training Phrases	$\rho$	0.819	0.231	0.406	0.464	<b>0.551</b>
	QWK	0.808	0.168	0.325	0.409	<b>0.479</b>
Responses	$\rho$	0.750	0.150	0.127	<b>0.235</b>	0.143
	QWK	0.715	0.083	0.105	<b>0.158</b>	0.094

# RQ1: How well do LLMs perform?

Artifact Dimensions	Metrics	Human-human	Human-GPT4				
			Zero-shot	Zero-shot	Few-shot	Few-shot	
			Basic	Rubric	Basic	Rubric	
Greet intent	$\rho$	0.850	0.339	0.641	<b>0.659</b>	0.646	} High agreement with human
	QWK	0.820	0.325	0.623	<b>0.698</b>	0.645	
Default Fallback intent	$\rho$	0.979	0.179	0.782	0.779	<b>0.816</b>	
	QWK	0.984	0.252	0.750	0.781	<b>0.797</b>	
Follow-up intents	$\rho$	0.839	0.133	0.217	0.203	<b>0.346</b>	Fair agreement
	QWK	0.805	0.154	0.244	0.230	<b>0.388</b>	
Training Phrases	$\rho$	0.819	0.231	0.406	0.464	<b>0.551</b>	Moderate agreement
	QWK	0.808	0.168	0.325	0.409	<b>0.479</b>	
Responses	$\rho$	0.750	0.150	0.127	<b>0.235</b>	0.143	Fair agreement
	QWK	0.715	0.083	0.105	<b>0.158</b>	0.094	



# RQ1: How well do LLMs perform?

Artifact Dimensions	Metrics	Human-human	Human-GPT4				
			Zero-shot	Zero-shot	Few-shot	Few-shot	
			Basic	Rubric	Basic	Rubric	
Greet intent	$\rho$	0.850	0.339	0.641	<b>0.659</b>	0.646	} High agreement with human
	QWK	0.820	0.325	0.623	<b>0.698</b>	0.645	
Default Fallback intent	$\rho$	0.979	0.179	0.782	0.779	<b>0.816</b>	
	QWK	0.984	0.252	0.750	0.781	<b>0.797</b>	
Follow-up intents	<p style="text-align: center;">LLM Challenges:</p> <p>1) carry out complex reasoning across multiple intents</p> <p>2) infer the logical progression of the conversation.</p>					Fair agreement	
Training Phrases						Moderate agreement	
Responses						Fair agreement	



# RQ2: Trade-offs among prompt strategies

Artifact Dimensions	Metrics	Human-human	Human-GPT4			
			Zero-shot	Zero-shot	Few-shot	Few-shot
			Basic	Rubric	Basic	Rubric
Greet intent	$\rho$	0.850	0.339	0.641	<b>0.659</b>	0.646
	QWK	0.820	0.325	0.623	<b>0.698</b>	0.645
Default Fallback intent	$\rho$	0.979	0.179	0.782	0.779	<b>0.816</b>
	QWK	0.984	0.252	0.750	0.781	<b>0.797</b>
Follow-up intents	$\rho$	0.839	0.133	0.217	0.203	<b>0.346</b>
	QWK	0.805	0.154	0.244	0.230	<b>0.388</b>
Training Phrases	$\rho$	0.819	0.231	0.406	0.464	<b>0.551</b>
	QWK	0.808	0.168	0.325	0.409	<b>0.479</b>
Responses	$\rho$	0.750	0.150	0.127	<b>0.235</b>	0.143
	QWK	0.715	0.083	0.105	<b>0.158</b>	0.094

# Human vs GPT-4 Scoring and Rationale

## Evaluation of **Greet Intent** Response of a Chatbot:

*“Hey, bro! My name is M&P game reccs, and you can ask me to start the quiz for my cracked game quiz to give you a board game rec, man!”*

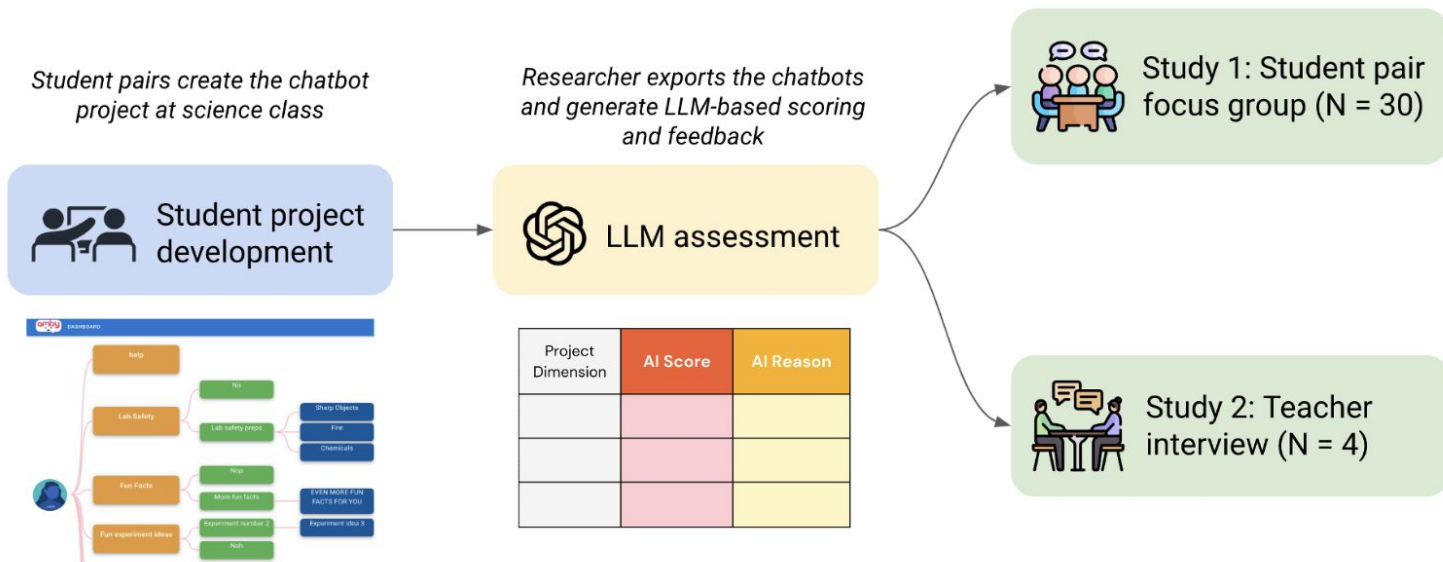
Evaluation	Score	Rationale
Human	4	<b>It sets user expectations properly because the main focus of the chatbot is the quiz which gives the user game recommendations.</b>
GPT-4: zero-shot-basic	2	The greeting does introduce the bot and its function, but <b>the language used is informal</b> and may not be understood by all users. <b>The message could be clearer and more professional.</b>
GPT-4: zero-shot-rubric	3	The greet intent is customized and introduces the chatbot as M&P game reccs. It also demonstrates its purpose by mentioning that it can start a quiz for game recommendations. However, <b>it doesn't set exact user expectations or clearly direct the user for next steps.</b>
GPT-4: few-shot-basic	3	The greeting is casual and friendly, but <b>it could be clearer in explaining the bot's purpose and how it can assist the user.</b>
GPT-4: few-shot-rubric	4	The greeting is customized, introduces the bot, demonstrates its purpose, and sets exact user expectations.

# Main Takeaways

- LLM-based artifact evaluation is task-dependent: strong on simple tasks, weaker on complex reasoning across intents.
- Few-shot prompting with examples improves grading accuracy.
- In cold-start cases (when no examples available), adding a rubric helps accuracy.

# A blurb about our follow-up study

- Manuscript in preparation: *Too Skeptical or Not Skeptical Enough? Middle School Student and Teacher Perceptions of LLM-Based Project Assessment*



Similarity		
Detailed reasoning and feedback	Both groups recognized the value of AI's ability to provide detailed feedback linked to specific rubric criteria and deliver potentially unbiased assessments.	
Concerns about accuracy for direct scoring	Both acknowledged the importance of accuracy and the need for AI to understand the specific learning context, including classroom content and individual student progress.	
Role of AI grading in classrooms	Both believed AI cannot take over teacher's grading. Teachers emphasized their responsibility of student grading and needs control over it. Students stated they trust teacher's grading more even if the score is lower. AI also needs to be transparent about the assessment process in order to be trustworthy.	
Importance of human factors	In project-based learning, both students and teachers care about creativity and individual differences, this part AI cannot supplement teacher's role.	
Value in Formative Feedback	Both believed AI for offering feedback during project development would be extremely beneficial, both for teachers to save time and for students to receive feedback early to improve their projects.	
Differences		
Aspect	Students Perspective	Teachers Perspective
Reactions on the chatbot assessment example	Mostly positive about the scores and reasoning	More critical about the generated content
Trust in AI-Generated Grades	Mixed trust: 20/30 trusted AI's judgment for grading, 12/30 preferred teacher grading. All agreed AI assessment is more trustworthy than assessment from a random adult outside of classroom.	Lower trust: All interviewed teachers expressed significant reservations about relying on AI for directly assigning grades.
Perceptions of AI Accuracy	Some viewed AI as more accurate and less error-prone than human, particularly in evaluating chatbots due to their shared AI nature.	More concerned about the accuracy, noting the importance of human judgment in assessing subjective aspects like creativity and student effort.
Practical Concerns	Concerned about potential threats to their own agency and autonomy, such as AI being overly directive in feedback.	Focused on potentially increased workload to safeguard the AI-generated content. Also noting logistical challenges related to implementation, data privacy and cross-platform integration.

Be on  
arxiv  
soon :)





This work is supported by National Science Foundation DRL-2048480.  
Thanks to all members of Project DIALOGS who contributed to this work.



## Full Paper Link



<https://tinyurl.com/chatbot-llm>

## Contact



Xiaoyi Tian  
xtian9@ncsu.edu

Tian, X., Mannekote, A., Solomon, C. E., Song, Y., Wise, C. F., Mcklin, T., ... & Israel, M. (2024). Examining LLM Prompting Strategies for Automatic Evaluation of Learner-Created Computational Artifacts. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 698-706).

# Additional slides

Table 6: Full Description of Conversational AI Artifact Evaluation Rubric.

Categories	Artifact Dimensions	1. Little to no evidence of approaching expectations	2. Approaching Expectations	3. Meets Expectations	4. Exceeds Expectations
Conversational design	Follow up intents	No follow-up intent	At least one follow-up intent OR most follow-up intents do not logically match with its parent intent OR they are unnecessary or repeated	Multiple logical follow-up intents AND Each follow-up intent is related to its parent intent mostly logically	All follow-up intents are logically related to main intent, numerous, and mutually exclusive
Conversational design	Greet intent	No customized greet response	At least one customized greet intent, however the purpose is not clear or actionable	At least one customized greet intent demonstrating its purpose. May not set exact user expectations: (“Ask me for song recommendations”, “hey im blah bot do you need any assistance on video games?” )	Effectively greet the user, introduce the chatbot, and demonstrate the purpose. AND Set exact user expectations (e.g., “I can talk about pop or hip hop music”) or clearly directs the user for next steps (e.g., “simply state ‘quiz me on math’”)
Conversational design	Default fallback intent	No customized fallback response	The response is customized, however it cannot not redirect the users (e.g., “I didn’t get that. Try it again.”)	The response is customized and can redirect the users (e.g., “I didn’t get that as I’m still learning. I’m more confident to talk about XYZ instead.”)	The agent has multiple varied, customized and meaningful responses that can redirect the users
AI Development	Training phrases	The amount of training phrases is limited (less than system requirement) OR Most of training phrases are random in the customized intents	The amount of training phrases meet the system requirement, but the content does not show enough linguistic variations (syntactically and lexically) within the intent or topic variations across different intents	Most training phrases are ample, cohesive and varied within the intent; also differ from those in other intents. They present variations in either syntactic structure or lexicon choices	The project contains consistently more varied training phrases than what the system requires, which can capture some edge cases. Training phrases are given and they are unique in both lexical and syntactic structure
AI Development	Responses	The responses are random in most of the customized intents	Most Responses (60%+) are provided either too long or too short, or lack of information or contains grammatical errors that impede user’s understanding If there are multiple responses, the content is not consistent enough to trigger similar user reactions Example: “Bad Romance by Lady Gaga” - not conversational	Most customized intents contain at least one response that is in proper length, logical, mostly free of grammatical errors, mostly mimic/display natural and conversational, may include some conversational markers.	Intents contain multiple logical, error-free responses OR The responses contain hints to keep the conversation going (e.g., “Alligators are dangerous animals... Now, do you want to learn about other animals? ) OR Utilize the conversational markers throughout the customized intents when appropriate