Examining LLM Prompting Strategies for Automatic Evaluation of Learner-Created Computational Artifacts

Xiaoyi Tian[1] (tianx@ufl.edu), Amogh Mannekote[1], Carly E. Solomon[1], Yukyeong Song[1], Christine Fry Wise[2], Tom Mcklin[2], Joanne Barret[1], Kristy Elizabeth Boyer[1], Maya Israel[1]

[1] *University of Florida*

[2] *The Findings Group*

Motivation

Evaluating learner projects and providing timely feedback is challenging because manual evaluation is time-consuming and resource-intensive.

Evaluating chatbots needs to consider both design and technical implementation and the logical flow of conversations.

Goal and Research Questions

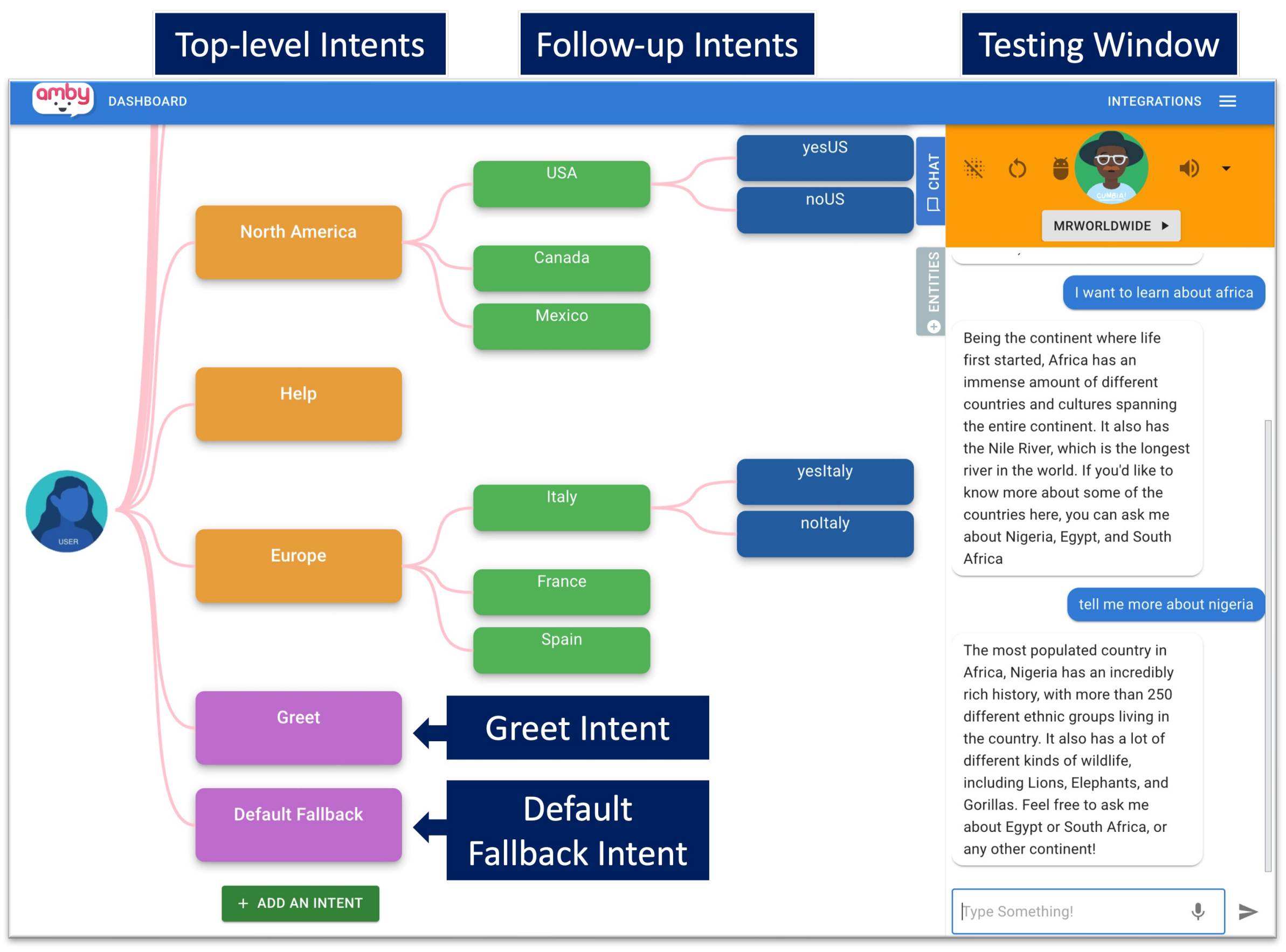
Investigate the capability of GPT-4 in automatically assessing student chatbot artifacts.

RQ1: How Do LLMs Perform in Assessing different Aspects of Computational Artifacts?

RQ2: What Are the Tradeoffs among Different Prompting Strategies?

Study Context

AMBY Chatbot Development Environment



I want to learn about africa

Being the continent where life first started, Africa has an immense amount of different countries and cultures spanning the entire continent. It also has the Nile River, which is the longest river in the world. If you'd like to know more about some of the countries here, you can ask me about Nigeria, Egypt, and South Africa

tell me more about nigeria

The most populated country in Africa, Nigeria has an incredibly rich history, with more than 250 different ethnic groups living in the country. It also has a lot of different kinds of wildlife, including Lions, Elephants, and Gorillas. Feel free to ask me about Egypt or South Africa, or any other continent!

Data: Program snapshots of 75 chatbots created mainly by middle school learners during an AI summer camp

Chatbot Artifact Evaluation Rubric

We focus on evaluating five artifact dimensions. Each dimension was rated on a 1-4 scale. Rubric overall Cohen's Kappa = 0.82.

Artifact Dimensions	Statement for Score of 3
Greet intent	At least one customized greet response demonstrating its purpose. May not set exact user expectations.
Default fallback intent	At least one customized default fallback response that can redirect the users.
Follow-up intents	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.
Training phrases	Most training phrases are ample, cohesive, and varied within the intent.
Responses	At least one response is of appropriate length, logical, conversational, and mostly free from grammatical errors.

Artifact Evaluation Implementation

Experimental Setup

LLM4Qual (github.com/msamogh/llm4qual) open-source framework for experiment

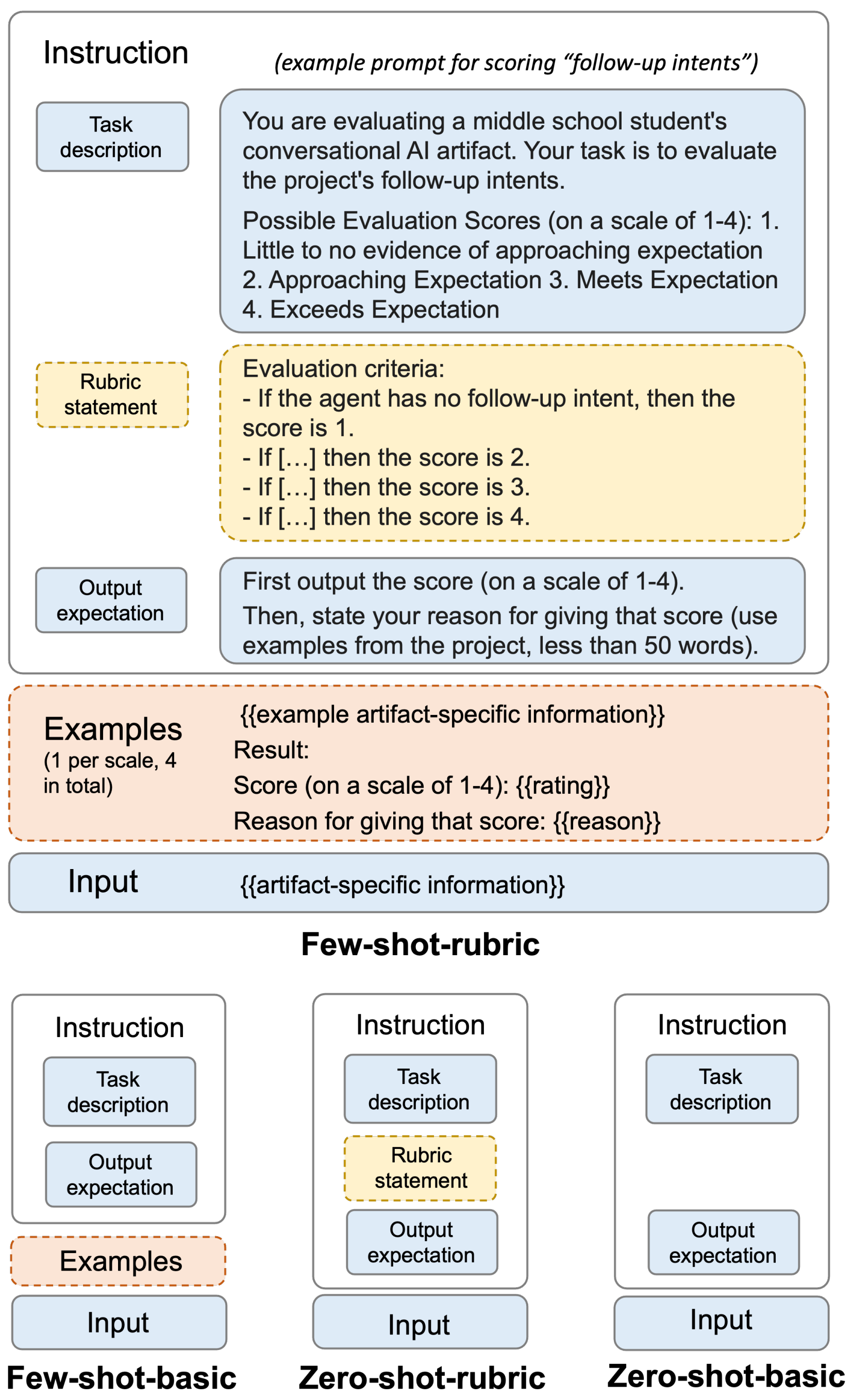
Four prompting strategies:

- zero-shot-basic
- zero-shot-rubric
- few-shot-basic
- few-shot-rubric

Evaluation Metrics

- Spearman correlation (ρ)
- Weighted Cohen's Kappa (QWK)

Prompt Template



(example prompt for scoring "follow-up intents")

You are evaluating a middle school student's conversational AI artifact. Your task is to evaluate the project's follow-up intents.

Possible Evaluation Scores (on a scale of 1-4): 1. Little to no evidence of approaching expectation 2. Approaching Expectation 3. Meets Expectation 4. Exceeds Expectation

Evaluation criteria:
- If the agent has no follow-up intent, then the score is 1.
- If […] then the score is 2.
- If […] then the score is 3.
- If […] then the score is 4.

First output the score (on a scale of 1-4).

Then, state your reason for giving that score (use examples from the project, less than 50 words).

{{example artifact-specific information}}

Result:

Score (on a scale of 1-4): {{rating}}

Reason for giving that score: {{reason}}

{{artifact-specific information}}

Figure 2. Prompt templates for four experiment conditions

Results

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

Table 2. Evaluation metrics of GPT4-generated scores with four prompting strategies and human-human agreement for five artifact dimensions

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

Table 3. Evaluation comparison of the Greet Intent Response: "*Hey, bro! My name is M&P game recs, and you can ask me to start the quiz for my cracked game quiz to give you a broad game rec, man!*"

Takeaways

- LLM-based evaluation performance is task-dependent. LLM performs well in evaluating isolated elements within a singular intent (like a greet response); however, their performance reduces when carrying out complex reasoning across multiple intents and inferring logical progression of the conversation (like follow-up intents and responses).
- Including a few contextual examples in the LLM prompts (few-shot strategy) can improve grading accuracy.
- When no contextual examples are available (cold-start problem), including a rubric statement can enhance accuracy.

Funded by NSF award DRL-2048480