Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI, the Arts and Creativity

Large Language Models for Human-AI Co-Creation of Robotic Dance Performances

Allegra De Filippo and Michela Milano
Department of Computer Science and Engineering, University of Bologna, Italy
allegra.defilippo@unibo.it, michela.milano@unibo.it

Abstract

This paper focuses on the potential of Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs), in the still unexplored domain of robotic dance creation. In particular, we assess whether a LLM (GPT-3.5 turbo) can create robotic dance choreographies, and we investigate if the feedback provided by human creators can improve the quality of the output. To this end, we design three prompt engineering techniques for robotic dance creation. In the prompts, we gradually introduce human knowledge through examples and feedback in natural language in order to explore the dynamics of *human-AI co-creation*. The experimental analysis shows that the capabilities of the LLM can be improved through human collaboration, by producing choreographies with a major artistic impact on the evaluation audience. The findings offer valuable insights into the interplay between human creativity and AI generative models, paving the way for enhanced collaborative frameworks in creative domains.

1 Introduction

Generative Artificial Intelligence (AI) has emerged as a prominent field of study, revolutionizing various domains, such as computer vision, natural language processing, and creative arts [Epstein *et al.*, 2023]. Generative AI models are characterized by the capability to autonomously generate novel and contextually relevant content. Leading this transition are Large Language Models (LLMs) [Naveed *et al.*, 2023; Wei *et al.*, 2023], and, in particular, consumer applications like ChatGPT[1] or DALL-E[2]. These generation models, commonly known as generative pre-trained transformers or GPT models, have been trained to comprehend both natural and formal language. Recent GPT versions, such as GPT-4 and GPT-3.5, are highly capable of producing text outputs based on users' requests, often referred to as *prompts* [Ekin, 2023; Wu and Hu, 2023].

In the context of LLMs, *prompting* refers to providing a specific input or instruction to the model to generate a desired output. Users can interact with these models by inputting a prompt, which can be a question, a statement, or any text that serves as an instruction for the model to generate a coherent response. LLMs can be prompted in various setups: 1) *Zero-Shot Prompting*, when LLMs are just required to answer to the query that is provided by the user; 2) *Few-Shot prompting,* when input-output demonstration pairs, showing correct answers, are shown to the model to generate the desired output; 3) *Reasoning in LLMs:* when LLMs are asked to generate answers to logical problems, task planning, with reasoning (*e.g.,* Chain-of-Thought for step-by-step reasoning [Wei *et al.*, 2023]).

Recent works have explored the potential of using LLMs in creative tasks, from narrative writing to poetry. An important aspect of this research is to investigate LLMs capabilities in terms of their ability to both generate creative content and assess whether an output is well artistically evaluated. Indeed, the creative process of generative AI is an open research field [Franceschelli and Musolesi, 2023]. Moreover, the rapid adoption of generative AI technologies by consumers is enabling a radically new way for people to interact with computing technology also in creative tasks [Deng and Lin, 2022]. Users are able to create specifications of the kinds of outputs they desire in different ways: natural language, sketches and gestures, and novel user interface controls. This enabling new forms of co-creativity while also posing new challenges of how to support users in creating outcome specifications that yield the desired results and are effective to use in different artistic domains [Mirowski *et al.*, 2023].

In this context, we focus on exploring the opportunities and challenges of creating new artistic outputs with generative systems in the domain of robotic dance creation, a still unexplored field of creative application for generative AI. We focus on how the capabilities of a generative model can be leveraged and enhanced through human-AI collaboration for producing robotic dance performances. Specifically, we design three prompting techniques as input to a LLM (GPT3.5-turbo) to create a robotic choreography for a NAO robot[3] as a *textual sequence of positions*. We explore the dynamics of human-AI co-creation by starting from designing a tailored

[1] https://openai.com/chatgpt
[2] https://openai.com/research/dall-e
[3] https://us.softbankrobotics.com/nao

zero-shot technique for this specific domain, and by gradually introducing human knowledge through examples and feedback in natural language. The ability to directly process natural language is a key element for enhancing human-AI co-creation. In our experimental analysis, we test the generated outputs (textual sequences of positions) by performing all of them with a humanoid NAO robot. The evaluation of the performances is then conducted with the help of a human audience unaware of the choreography creation processes, and using a state-of-the-art evaluation scheme. The experimental analysis shows that the capabilities of the considered LLM can be improved through human co-creation, by producing choreographies with a major artistic impact.

To summarize, this paper provides the following contributions: *(1)* We assess the effectiveness of LLMs for creating robotic dance choreographies; *(2)* We design three prompting techniques tailored for an unexplored artistic domain, by focusing on human-AI co-creation; *(3)* We evaluate and analyse the artistic impact of the generated outputs for the different prompting techniques. From our perspective, this work findings offer valuable insights into the interplay between human creativity and AI generative models in a still unexplored artistic domain, by paving the way for enhanced collaborative frameworks in creative domains, by holding the promise of pushing the boundaries of what can be achieved.

2 Background

Robotic Dance Creation and Evaluation In recent years, many researchers proposed methods to automate partial aspects of dance, from dance notation to choreography, and from dance capture to dance generation [Sagasti, 2019; Joshi and Chakrabarty, 2021]. In the artistic domain of dance, the physical movement is a key factor. For this reason, the use of robots is continually expanding thanks to their humanoid shape. Many works have studied and implemented sophisticated systems for robotic dances, thanks to improvements in mechanics and control [Ramos *et al.*, 2015; Shinozaki *et al.*, 2007; Shinozaki *et al.*, 2008; Aucouturier *et al.*, 2008] . Due to the recent expansion of this field, some recent works started focusing on the definition of a common setting for the aesthetic evaluation of these artistic outputs [Oliveira *et al.*, 2012; De Filippo *et al.*, 2022b; De Filippo *et al.*, 2023]. However, all these works disregard the actual degree of human collaboration and intervention in the creative process which, on the contrary, is an important feature to be taken into account [Hong and Curran, 2019]. For these reasons, starting from an AI technique that is suitable for human collaboration (*i.e.,* Large Language Models), in our work we analyze how different prompting techniques for robotic dance creation can affect the aesthetic evaluation of a generated output by gradually increasing the human knowledge through examples and human feedback.

Generative AI and LLMs Generative AI refers to systems that have the capability to produce new, original content [Sætra, 2023]. These systems are designed to autonomously generate outputs, often mimicking the patterns and styles learned from large datasets during their training. Large Language Models (LLMs) are a specific category of generative AI that focuses on language-related tasks, and trained on vast amounts of textual data, they can understand, generate, and manipulate human-like language [Fui-Hoon Nah *et al.*, 2023]. These models operate through learned patterns from massive datasets, enabling them to generate creative and coherent outputs based on given inputs, often referred to as prompts [Zamfirescu-Pereira *et al.*, 2023]. Effective prompts empower users to leverage the powerful capabilities of LLMs, obtaining accurate and relevant responses that enhance capabilities [White *et al.*, 2023]. Due to their substantial computational resources required for re-training such models [De Filippo *et al.*, 2022a], prompting [Han *et al.*, 2021] emerges as a valuable technique for adapting pre-trained models without incurring the expenses associated with a full fine-tuning procedure [De Filippo *et al.*, 2019]. ChatGPT is a well known application based on generative pre-trained Transformer GPT that uses the Reinforcement Learning with Human Feedback approach [Stiennon *et al.*, 2020]. This holds promise as an effective means to align Large Language Models with human intent [Ouyang *et al.*, 2022], and for tackling novel tasks not originally targeted during training, through the use of specially tailored prompts. Despite the significance of prompt engineering [Saravia, 2022; Ekin, 2023], there remains a research gap regarding the impact of prompt engineering on creativity tasks [Chakrabarty *et al.*, 2023a], and how human knowledge and collaboration can enhance their creativity [Franceschelli and Musolesi, 2023]. Therefore, the main aim of our paper is to investigate the influence of prompt engineering with a progressive insertion of human knowledge in form of examples and feedback in natural language, by grounding our analysis on the creative context of robotic dance creation. To the best of our knowledge, this is a still unexplored field for LLMs generation for creativity and human collaboration.

Human-AI Co-Creation Recent progress in AI, especially in advanced models like generative ones, has brought attention to how AI can assist in creative collaborations between humans and machines [Fui-Hoon Nah *et al.*, 2023]. In a co-creative context, for example, AI might help improve a user's drawing [El-Zanfaly *et al.*, 2023], write sentences of a story alongside a human [Chakrabarty *et al.*, 2023b], or fill in missing parts of a user's music composition [Huang *et al.*, 2020]. Recently, this collaborative way of creation is especially focused on artistic fields [Franceschelli and Musolesi, 2023], and the possibility to directly communicate with users through natural language and the rapid diffusion of consumer applications like ChatGPT are fueling the growth of human-AI co-creation applications. However, some artistic fields are still unexplored. Therefore, with the aim of starting to bridge the gap in this domain, in this work we focus on human-AI collaboration through LLMs for robotic dance creation.

3 Prompt Engineering for LLMs

Prompt engineering is essential in guiding the generation of desired responses and output, and it gradually emerges as a popular paradigm to control the behavior of LLMs, since it can effectively adapt a pre-trained model to downstream tasks in either zero-shot or few-shot style. Prompts facilitate communication between users and LLMs, by providing

guidance to ensure the generation of outputs aligned with the user's intent. As a result, well-engineered prompts greatly improve the efficacy and appropriateness of LLMs outputs [Zamfirescu-Pereira *et al.*, 2023].

3.1 Prompts Design

In this work, we designed a prompt to enable a LLM to generate a robotic dance choreography. In particular, we started by a *zero-shot* prompt, where we just ask the LLM to generate the choreography, then we enrich it by using examples and human feedback. In our implementation, we follow some essential components and guidelines to design a well-made prompt [Ekin, 2023; Chen *et al.*, 2023]. In particular:

- *Giving instructions.* A comprehensive description is imperative to elicit more precise and relevant outputs [Zhou *et al.*, 2023; OpenAI, 2024]. In our case, we create a tailored prompt with a set of requirements in form of instructions, such as *"Create a choreography as a sequence of positions for a NAO robot. Create the choreography for a rock song."*
- *To be clear and precise.* This involves formulating prompts that are unambiguous, which can guide the model toward generating the desired output [Aljanabi *et al.*, 2023; OpenAI, 2024]. Our prompt is structured in short and clear sentences, such as *"Write your output as a list of positions. Positions can be repeated."*
- *Role-prompting.* It involves giving the model a specific role to play, such as a knowledgeable expert [Wang *et al.*, 2023b; OpenAI, 2024], for guiding the model's responses in alignment with the desired output. We design our prompt by giving a role to the model, such as *"You are a robotic dance choreographer."*
- *Use of quotes and delimiters.* This technique is particularly useful when dealing with complex prompts that include multiple components, which makes the model understand one's instructions better [OpenAI, 2024]. We delimit our prompt in lists of position classes, to better organize our prompt: *"Mandatory positions:* $[Mandatory\_sit, ..., Mandatory\_zero]$*."* More details about positions will be provided in Section 4.

3.2 LLM Parameter Setting and API

The output of LLMs can be affected by various hyperparameters, whose setting plays a crucial role in the generation of outputs [OpenAI, 2024; Wang *et al.*, 2023a]. We briefly describe the configurable hyper-parameters that we used in our setting: *temperature* and *top-p*. The temperature parameter, with values ranging between $[0, 1]$, controls the randomness of the generated output: a lower temperature leads to more deterministic outputs. The top-p parameter, on the other hand, controls the nucleus sampling, which is a method to add randomness to the model's output, and the values range is the same. Adjusting these parameters can significantly affect the quality and diversity of the model's outputs, making them essential tools in prompt engineering [Lee *et al.*, 2023; Lee *et al.*, 2023]. In this work, we used OpenAI API[4] and GPT3.5-turbo, in particular *gpt-3.5-turbo-1106* model[5] that is the latest version of GPT3.5 and it also allows to set parameters for obtaining more reproducible outputs. Indeed, we set two further parameters: *n* and *seed*. The n parameter is related to the possibility of generating multiple outputs given a single prompt (we set this parameter to 1). Fixing the seed parameter allows to control that repeated requests with the same parameters should return the same (or similar) result.

We conducted a first exploratory analysis of different parameter configurations, in order to analyse their impact on the generated outputs. Finally, we configure our experimental setting by fixing *temperature* to 0.7 and *top-p* to 0.8, as recommended for achieving a good balance between creativity and respect of requirements [OpenAI, 2023], and by leaving all the remaining parameters of *gpt-3.5-turbo-1106* to their default settings.

4 Methodology

As illustrated in Figure 1, our approach is based on three macro-phases: 1) artistic domain setting definition, 2) robotic choreography creation based on different prompting techniques with increasing degree of human co-creation, 3) artistic evaluation phase of the generated outputs.

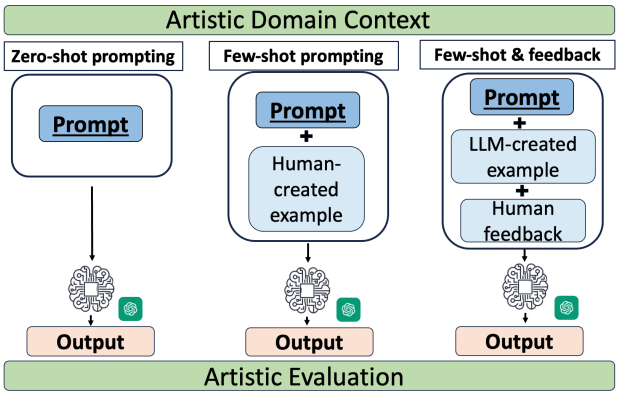

Figure 1: Prompting techniques for robotic dance creation

In details, we first define and formalize the domain setting of robotic choreography creation through an accurate setting description in terms of robot positions and artistic/technical constraints to be respected. Then, we design three prompting techniques as input to a LLM (GPT3.5-turbo). The aim is to create a robotic choreography, as a textual sequence of positions, suitable for a NAO robot. Based on the three different prompting techniques, we explore the dynamics of human-AI co-creation by gradually introducing human knowledge through examples and feedback in natural language. Our techniques are: 1) Zero-Shot, 2) Few-Shot, 3) Few-Shot & Human Feedback. In our experimental analysis, we test the generated outputs (sequences of positions) by performing all of them with a NAO robot. The evaluation of the performances is then conducted with the help of a human audience unaware of the choreography creation processes, and using

[4]https://platform.openai.com/docs

[5]https://platform.openai.com/docs/models/gpt-3-5

a state-of-the-art evaluation scheme. By following the same structure and the same steps, our approach can be extended to different artistic domains.

4.1 Robotic Dance Domain Setting

As previously stated, the aim of this work is to generate a choreography (expressed as a sequence of positions) to be tested on a simulated NAO, a humanoid robot developed by SoftBank Robotics[6]. In our setting, we start with a state-of-the-art *problem description* [De Filippo *et al.*, 2023], which consists of the definition of a set of *positions* suitable for the NAO robot, split into mandatory and intermediate positions (*e.g.,* sit position, stand position, etc.). In order to be passed as input of a LLM in natural language, we propose clear and explicative names for each position. Moreover, we define a set of representative music genres [Sturm, 2012; Scaringella *et al.*, 2006] (*i.e.,* folk, electronic, classical, rock and latin).

Based on recent works in this application domain that represent choreographies as sequences of basic positions for robotic dances [De Filippo *et al.*, 2023; Oliveira *et al.*, 2012; Liu *et al.*, 2020; Wang, 2022], we list a set of constraints to be respected in the choreography creation: (1) each choreography must start with a requested *initial* position and must end in a requested *final* position; (2) the total duration must be of 2 minutes; (3) each choreography must contain at least each mandatory position; (4) each choreography must contain at least 5 different intermediate positions; (5) each choreography must be associated with a music track. The requirements of our domain setting are summarized in Table 1.

list_mandatory	[Mandatory_sit, Mandatory_wipe_forehead, Mandatory_hello, Mandatory_stand, Mandatory_zero]
list_intermediate	[rotation_handgun_object, right_arm_rotation, double_movement_rotation_of _arms, arms_opening, union_arms, move_forward, move_backward, diagonal_left, diagonal_right, rotation_foot_left_leg, rotation_foot_right_leg, play_guitar, arms_dance, birthday_dance, sprinkler_dance, workout_legs_and_arms, superman]
initial and final	[INITIAL_stand_init, FINAL_crouch]
music genres	[Folk, Electronic, Classical, Rock, Latin]

Table 1: Requirements of positions and music in our domain setting

4.2 Prompt Engineering for Robotic Dance Creation

We describe the design and the implementation of our three prompting techniques. The generated choreographies and the source code are available in our private repository.[7]

Zero-shot Prompting First, we design a basic Zero-shot prompt, as depicted in Figure 2, and we define role, context, instructions and constraints, by following the design rules illustrated in Section 3. The keywords *[list_mandatory]* and *[list_intermediate]* are referred to the respective lists of positions and *[music_genre]* refers to the specific music genre (among those on the list) and the fixed time duration of 2 minutes. Based on this prompt, the LLM creates a choreography as a sequence of positions that follows the instructions. An example of the output is provided in Figure 5 in Section 5.1.

Prompt

Role and context: "You are a robotic dance choreographer. You will be provided with a set of constraints to be respected for creating the most suitable sequence of positions (i.e., a choreography) for a NAO robot. You will be provided with a musical genre and a time duration for the choreography. Write your output as a list of positions."

Instructions and constraints: "Create a choreography as a sequence of positions for a NAO robot, by respecting the following constraints: -Start position: [INITIAL_stand_init]. -End position: [FINAL_crouch]. -Mandatory positions: [list_mandatory]. -Intermediate positions: [list_intermediate]. - Use all the mandatory positions. - Use at least 5 intermediate positions. - Positions can be repeated. - Shuffle the order of mandatory and intermediate positions. - Create the choreography for [music_genre]. After creating the choreography, explain your choices as a list of motivations for the whole choreography and its association with the given music."

Figure 2: Zero-Shot Prompting

Few-shot Prompting Next, we improve our first technique by adding a human-created output example. As shown in Figure 3, we start with the *zero-shot* prompt and we juxtapose an example choreography created by a human choreographer aware of the artistic domain setting[8]. In this case, the resulting choreography (see an example again in Figure 5) is supposed to also exploit the knowledge explicitly provided by the choreographer.

Human example: "1.INITIAL_stand_init, 2.Mandatory_zero, 3.union_arms, 4.birthday_dance, 5.move_backward, 6.Mandatory_sit, 7.play_guitar, 8.move_forward, 9.move_backward, 10.Mandatory_stand, 11.Mandatory_wipe_forehead, 12.workout_legs_and_arms, 13.sprinkler_dance, 14.Superman, 15.play_guitar, 16.diagonal_left, 17.move_backward, 18.Mandatory_sit, 19.Mandatory_stand, 20.Mandatory_hello, 21.Mandatory_hello, 22.superman, 23.FINAL_crouch."

Figure 3: Few-Shot Prompting

Few-shot & Human Feedback Prompting Finally, the third technique, depicted in Figure 4, further extends the *few-shot* prompt with an explicit feedback provided by a human choreographer based on the artistic domain knowledge. It is important to underline that, in this technique, we add to the basic prompt *the output of the previous (Few-Shot) technique* as example, and we also add human feedback. In this case, our aim is to refine the output of the LLM through the feedback of a human choreographer, in order to deepen the degree of co-creation. This technique is intended to produce output that is more tailored in terms of movements more coherent with the music genre and with a major artistic impact.

[6]https://www.softbankrobotics.com

[7]https://anonymous.4open.science/r/LLMsChoreography-601F

[8]Due to the maximum length of prompt fixed by OpenAI API, we use a single example, but our prompt can be easily extended.

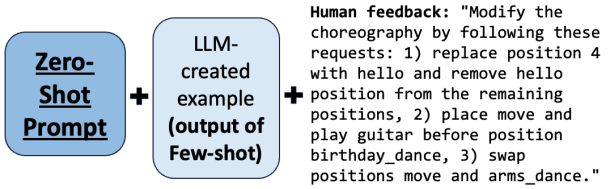

Figure 4: Few-shot & Human Feedback Prompting

4.3 Artistic Evaluation Phase

The evaluation phase is conducted to investigate the reactions and perceptions of the audience after seeing the robotic performances created. The methodological tools used for data collection are participant observations and questionnaires. In particular, we define a survey[9] based on an ad-hoc questionnaire proposed in the state of the art [De Filippo *et al.*, 2023; Oliveira *et al.*, 2012] to evaluate the robotic choreographies generated through the different techniques.

In our setting, given a randomly selected music genre, each participant evaluated *three* choreographies, one for each prompting technique, for that genre. Next, the same process was repeated for a second (randomly selected, again) genre. So, each participant evaluated *six* choreographies in total. Each questionnaire is composed by 2 macro steps (one per different music genre), and each step is composed by three video demos of the created robotic choreographies and a list of questions, one for each evaluation target. Each participant anonymously vote the proposed choreographies, providing a score for all targets on a Likert scale (from 1 to 5). The evaluation targets are: (1) Storytelling; (2) Rhythm; (3) Movement Technique; (4) Public Involvement; (5) Space Use; (6) Human Characterization; (7) Human Reproducibility. For each target, we propose a specific question to the user.

5 Experimental Analysis

Our experimental analysis aimed to answer to the following research questions (RQs):

- **RQ1:** Do LLMs effectively create coherent robotic dance performances?
- **RQ2:** Does the adoption of a co-creation strategy (through feedback and examples) allow the audience perceive an aesthetic improvement in the resulting choreographies?
- **RQ3:** Does the music genre influence the evaluation of the audience on the different choreographies?

5.1 Experimental Setting

Results Collection and Audience. We conduct our experiment by following a *whithin-subjects* protocol. In particular, each user is exposed to a single subset of music genres and evaluated all the different prompting techniques. This allows to evaluate the differences among all the subsets of choreographies without any conditioning in the evaluation phase. Of course, the audience was not aware of the specific prompting technique was assigned to [Charness *et al.*, 2012]. To avoid decision fatigue [Pignatiello *et al.*, 2020], questions order is randomized in each questionnaire. To recruit the audience, we used Prolific[10] platform and the participants were prescreened by following these requirements: (1) Approval rate on the platform: 95–100; (2) Fluent languages: English; (3) Previous submissions: 5–10000; (4) Countries: France, Germany, Ireland, Italy, Spain, UK and USA.

Dataset Construction. For our dataset construction and analysis, we define independent variables in order to observe and measure their effects on our dependent variables. In details, our *independent variables* are 1) the creation process of the choreography (e.g., Zero-Shot) and 2) the music genre; the *dependent variables* are the evaluation targets that we want to measure. To answer RQ1, we first carried out a qualitative and quantitative analysis of the different choreographies, by comparing them in terms of requirements satisfactions and coherence with the given prompt, based on the generation technique. Then, to answer RQ2 and RQ3, we analyse the results of the artistic evaluation by comparing user scores in two different analyses: first, we compare the average scores obtained by the choreographies on varying of different prompting techniques, focusing on the specific evaluation targets; second, we compared the average scores for different music genres, again focusing on specific targets. The choreographies and the results we collected are available in our private repository.[7]

5.2 RQ1: Output Analysis

In Figure 5, we illustrate an example of the two generated outputs by using two different prompting techniques for the electronic music genre.

ZERO-SHOT	FEW-SHOT & FEEDBACK
1. INITIAL_stand_init	1. INITIAL_stand_init,
2. arms_dance	2. move_forward,
3. Mandatory_wipe_forehead	3. Mandatory_hello,
4. play_guitar	4. play_guitar,
5. Mandatory_sit	5. move_forward,
6. rotation_handgun_object	6. arms_dance,
7. move_forward	7. movement_rotation_of_arms
8. birthday_dance	8. move_backward,
9. Mandatory_hello	9. rotation_foot_left_leg,
10. rotation_foot_left_leg	10. Mandatory_sit,
11. arms_opening	11. rotation_handgun_object,
12. move_backward	12. Mandatory_stand,
13. workout_legs_and_arms	13. birthday_dance,
14. movement_rotation_of_arms	14. superman,
15. superman	15. Mandatory_wipe_forehead,
16. union_arms	16. sprinkler_dance,
17. diagonal_left	17. move_forward,
18. Mandatory_zero	18. move_forward,
19. right_arm_rotation	19. diagonal_right,
20. sprinkler_dance	20. union_arms,
21. FINAL_crouch	21. move_backward,
	22. Mandatory_zero,
	23. rotation_foot_right_leg,
	24. diagonal_left,
	25. FINAL_crouch

Figure 5: Different outputs generated for **electronic music genre**

We can notice macro-difference in the generated outputs of two different techniques. Therefore, we deepen our anal-

[9]https://forms.gle/gWCSaYGTpry61e8AA

[10]https://www.prolific.com

ysis by considering different metrics: the requirements satisfaction, and the presence of patterns in the positions sequence, that are typical in dance choreographies [Brown *et al.*, 2006]. In details (see Table 2), we consider for each

Prompting Technique	avg tot positions	avg position repetition	avg diff positions	Position req satisfied (%)	Time req satisfied (%)
Zero-Shot	22.2	0	22.2	0.2	0.4
Few-Shot	24.6	3	21.6	1	1
Few-Shot & Feedback	25.4	4.4	21	1	1

Table 2: Analysis of the outputs generated with the three techniques

group of choreographies created with the same technique: (1) the percentage of requirements satisfaction in terms of initial, final, mandatory and intermediate positions; (2) the percentage of requirement satisfaction in terms of total time duration of 2 minutes (with a precision interval of $\pm$ 5 seconds); (3) the average number of total positions; (4) the average number of position repetitions; and (5) the average number of different unique positions. As shown in Table 2, choreographies created through zero-shot prompting do not satisfy all the requirements in terms of positions and time duration, while few-shot prompting is able to satisfy the requirements, and also to generate patterns of positions through repetitions. These results generally confirmed our conjecture, showing that human feedback and human examples play a key role in the co-creation of artistic output based on LLMs. Indeed, zero-shot prompting is more effective when the LLM *already* holds the knowledge to fulfill the specific request of the user. On the contrary, when such knowledge is not encoded in the LLM, as it happens in our scenario, human examples are fundamental to empower the capabilities of LLMs in novel domain of applications, such as the artistic one.

These results suggest that *LLMs can effectively create coherent robotic dance performances through Few-Shot prompting by using human-generated examples and feedback*. The model is able to learn the suitable average number of total positions to respect the duration requirement, in Few-Shot settings. Moreover, based on the increasing introduction of human collaboration, the model is also able to learn patterns of movement repetitions.

5.3 RQ2: Generative Technique Comparison

To address RQ2, we analyse the results of the artistic evaluation, by comparing the average ratings received for the choreographies generated with the three different prompting techniques (our independent variables). Since the requirement of normal distribution is not met for these samples, we used a Kruskal-Wallis test [McKight and Najab, 2010] for the three independent samples, then for each significant results (i.e., p-value < 0.05), we used a two-sided Mann-Whitney U rank test [McKnight and Najab, 2010] on each pair of independent samples of collected scores. We show in Figure 6 the average scores and the standard deviation for each evaluation target and for each prompting technique.

The results showed that, for almost every evaluation target (with the exception of Human Reproducibility), the introduction of human feedback and examples significantly improves the perceived quality of the choreographies w.r.t. Zero-shot prompting. Moreover, in four out of seven targets (Storytelling, Rhythm, Public, and Human Characterization, highlighted with a *double blue star*) the outcomes are even more significant. Indeed, the results show that the injection of a *human feedback* in the prompts significantly improves the quality of the choreographies w.r.t. those generated through *both* zero-shot and few-shot prompting.

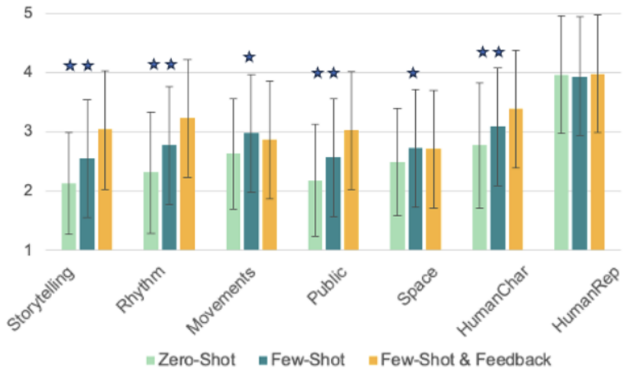

Figure 6: Average scores and deviation standard per evaluation target, for choreographies created with the three prompting techniques.

This suggested that: (1) participants generally have statistically significant preference for Few-Shot and Few-Shot & Feedback techniques for all the targets w.r.t. Zero-Shot technique; (2) participants positively evaluated the impact of human feedback in terms of storytelling of the performance, rhythmic coherence with the music, impact on the public, and human characterization of the robot performer; (3) participants significantly preferred choreographies created with Few-Shot and Feedback techniques based on the movement difficulty and the use of space, without significant preferences between these two techniques that are perceived as similar. This might be explained by the fact that human choreographers are used to work with real environments (the actual stage with its confined spaces, the position of the dancers and audience, etc) and thus might not have the right impact on the artificial space where the robot is dancing. Moreover, we used a finite set of positions, and this is in line with what we expected since the human expertise can only act on the order of positions. (4) As for the Human Reproducibility target, we asked if the choreography can be easily reproduced by a human performer, and the results showed that evaluators had no significant preferences among the different techniques. This can be a consequence of the same performer (*i.e.,* the NAO robot) used for all the choreographies. *These results clearly suggested that the audience significantly perceives an improvements in choreographies created through human collaboration, in particular for the techniques with human feedback*. Techniques with human feedback are able to produce outputs with a major impact on the audience in terms of coherence with the rhythm, storytelling capability of the whole choreography, impact of the performance on the public and also its perceived human characterization.

5.4 RQ3: Music Genre Comparison

The previous test revealed that there is statistically significant correlation between rhythmic coherence and music genre. Therefore, to address RQ3 and to better deepen this analysis, we further split the data based on the music genre. We used again the same test procedure on the independent samples of collected score. All statistically significant values are highlighted with (*) for gap between Zero-Shot techniques and the others (i.e., no significant preferences between Few-Shot and Few&Feedback techniques); and with (**) for significant gap among all the techniques, with an absolute preference for the Few&Feedback technique.

Music Genre	Ev Target	Zero-Shot		Few-Shot		Few&Feedback	
		avg	std	avg	std	avg	std
Folk	Storytelling	2.07	1.05	**2.63** (*)	0.99	**2.87** (*)	1.19
	Rhythm	2.27	0.98	**2.70** (*)	0.95	**3.03** (*)	1.24
	Movements	1.80	0.92	**3.07** (*)	0.87	**2.87** (*)	0.89
	Public	2.03	0.76	**2.40** (*)	0.89	**2.67** (*)	1.24
	Space	2.01	0.94	**2.63** (*)	0.67	**2.67** (*)	1.03
	Human Char	3.77	1.17	**3.01** (*)	0.94	**3.00** (*)	1.21
	Human Rep	2.63	0.99	4.03	0.87	3.97	0.99
Electronic	Storytelling	2.07	0.78	2.43	0.86	2.97	1.30
	Rhythm	2.17	1.02	2.60	0.81	**3.27** (**)	1.14
	Movements	2.57	0.89	2.73	0.94	2.76	1.00
	Public	2.33	0.92	2.56	0.89	**3.23** (**)	1.35
	Space	2.57	0.97	2.60	0.89	2.70	0.70
	Human Char	3.00	0.96	3.00	0.95	3.03	1.16
	Human Rep	4.07	0.78	3.93	0.86	4.03	0.61
Rock	Storytelling	2.06	0.91	2.26	0.86	**2.65** (**)	1.15
	Rhythm	2.17	1.05	**2.72** (*)	1.02	**2.93** (*)	1.11
	Movements	2.46	1.10	2.70	0.91	2.76	0.89
	Public	2.30	0.98	2.40	0.81	2.65	1.18
	Space	2.33	0.95	2.61	0.93	2.43	1.10
	Human Char	3.01	0.99	3.00	1.01	3.01	1.11
	Human Rep	3.66	1.09	3.81	0.88	3.6	1.13
Latin	Storytelling	2.15	0.64	2.90	0.99	**3.43** (**)	1.04
	Rhythm	2.73	0.94	3.13	0.93	**3.66** (**)	1.15
	Movements	2.91	0.71	3.26	0.69	3.03	1.09
	Public	2.30	0.87	3.03	1.06	**3.45** (**)	0.93
	Space	2.81	0.66	3.03	0.81	3.00	0.85
	Human Char	2.86	1.13	3.13	0.97	**3.63** (**)	1.32
	Human Rep	4.13	0.93	4.00	0.98	4.13	1.01
Classical	Storytelling	2.26	0.91	2.50	0.97	**3.20** (**)	1.21
	Rhythm	2.23	1.07	2.70	0.86	**3.23** (**)	1.27
	Movements	2.86	0.97	3.13	0.89	2.90	0.95
	Public	2.16	0.98	**2.73** (*)	0.97	**3.06** (*)	1.25
	Space	2.70	0.95	2.73	0.94	2.70	0.91
	Human Char	2.87	1.22	3.06	1.08	**3.53** (**)	1.25
	Human Rep	4.11	0.94	3.96	1.03	4.16	1.01

Table 3: Average scores and standard deviation per evaluation target based on music genre, for all the prompting techniques.

The test results in Table 3 showed that folk is the music genre with a statistically significant preference in terms of scores for almost all the targets for the choreographies created with the two Few-Shot techniques w.r.t. the Zero-Shot technique. For this music genre, the human feedback is not able to produce a significant improvement for a major impact on the audience. For all the remaining genres, the significant gaps are confirmed with a general preference for choreographies created with Few&Feedback for rhythmic coherence, storytelling, public and human characterization.

This third set of experiments confirms that similar score values can be observed for these targets: Movements (related to the technique and fluidity of movements), Space (the use of the space by the robotic performer), and Human Reproducibility (the possibility to reproduce the performance by a human performer). Average preferences connected to these targets are similar regardless of music genre. This can be explained by the common initial setting that limits both the movements choice and the degree of creativity allowed for public involvement, by confirming the trend observed in the second experiment. *Also in this setting, these results suggested that the audience significantly perceives an improvements in choreographies created through human collaboration.* In general, these results confirm that *for all the selected music genres* a significant improvement in the choreography impact is perceived by adding human co-creation techniques.

6 Discussion and Conclusions

In this work, we focused on the potential of LLMs, in conjunction with human creators, in the still unexplored domain of robotic dance creation. We investigate how the capabilities of a LLM (GPT-3.5 turbo) can be enhanced through human collaboration, by proposing a methodological approach to design, implement and evaluate different prompting techniques in this artistic domain. In our results, we assess the effectiveness of the LLM for creating robotic dance choreographies. Moreover, we show and analyze that the capabilities of the LLM are significantly improved through human co-creation, by producing choreographies with a major artistic impact on the evaluation audience.

To sum up, this work showed that technologies such as LLMs empower artists to collaborate with AI systems, facilitating a synergistic relationship that enhances the choreographic process. Moreover, the proposed method is designed to be easily extended both for different LLMs and for different creative domains. We plan to generalize our method as future work. As a further remark, we want to emphasize that one crucial aspect contributing to the success of this collaboration is the role of *explanation* provided by AI systems. As human choreographers engage in the co-creation process, the ability of AI to transparently articulate its decision-making and generative processes becomes essential. This explanatory capability not only clarifies the AI's creative contributions but also fosters a deeper understanding between human and machine collaborators. By comprehending the AI's thought processes, artists can adapt and guide the generative system more effectively, leading to a more symbiotic and harmonious co-creation of artistic choreographies.

While an example of such *explanations* were already generated through our prompts (e.g., *"The play_guitar position is chosen to synchronize with the electronic music's rhythm and create an engaging visual effect. The use of move_forward and move_backward positions adds dynamism and movement to the choreography, enhancing the visual appeal [...]"*), this analysis is left as future work: the transparency facilitated by explanations in LLMs and generative AI can enhance the creative dialogue, in a dynamic partnership that transcends traditional boundaries and opens up new horizons for artistic expression.

Acknowledgements

This work was funded by the PNRR - M4C2 (Investimento 1.3) Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Commission (NextGeneration EU program).

References

[Aljanabi *et al.*, 2023] Mohammad Aljanabi, Mohanad Ghazi Yaseen, Ahmed Hussein Ali, and Mostafa Abdulghafoor Mohammed. Prompt engineering: Guiding the way to effective large language models. *Iraqi Journal for Computer Science and Mathematics*, 4(4):151–155, 2023.

[Aucouturier *et al.*, 2008] Jean-Julien Aucouturier, Katsushi Ikeuchi, Hirohisa Hirukawa, Shin'ichiro Nakaoka, Takaaki Shiratori, Shunsuke Kudoh, Fumio Kanehiro, Tetsuya Ogata, Hideki Kozima, Hiroshi G Okuno, et al. Cheek to chip: Dancing robots and ai's future. *IEEE Intelligent Systems*, 23(2):74–84, 2008.

[Brown *et al.*, 2006] Steven Brown, Michael J Martinez, and Lawrence M Parsons. The neural basis of human dance. *Cerebral cortex*, 16(8):1157–1167, 2006.

[Chakrabarty *et al.*, 2023a] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity, 2023.

[Chakrabarty *et al.*, 2023b] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*, 2023.

[Charness *et al.*, 2012] Gary Charness, Uri Gneezy, and Michael A Kuhn. Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, 81(1):1–8, 2012.

[Chen *et al.*, 2023] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.

[De Filippo *et al.*, 2019] Allegra De Filippo, Michele Lombardi, Michela Milano, et al. How to tame your anticipatory algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1071–1077, 2019.

[De Filippo *et al.*, 2022a] Allegra De Filippo, Andrea Borghesi, Andrea Boscarino, and Michela Milano. Hada: An automated tool for hardware dimensioning of ai applications. *Knowledge-Based Systems*, 251:109199, 2022.

[De Filippo *et al.*, 2022b] Allegra De Filippo, Paola Mello, and Michela Milano. Do you like dancing robots? ai can tell you why. In *PAIS 2022*, pages 45–58. IOS Press, 2022.

[De Filippo *et al.*, 2023] Allegra De Filippo, Luca Giuliani, Eleonora Mancini, Andrea Borghesi, Paola Mello, Michela Milano, et al. Towards symbiotic creativity: A methodological approach to compare human and ai robotic dance creations. In *Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5806–5814, 2023.

[Deng and Lin, 2022] Jianyang Deng and Yijia Lin. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83, 2022.

[Ekin, 2023] Sabit Ekin. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.

[El-Zanfaly *et al.*, 2023] Dina El-Zanfaly, Yiwei Huang, and Yanwen Dong. Sand-in-the-loop: Investigating embodied co-creation for shared understandings of generative ai. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*, pages 256–260, 2023.

[Epstein *et al.*, 2023] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.

[Franceschelli and Musolesi, 2023] Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*, 2023.

[Fui-Hoon Nah *et al.*, 2023] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. Generative ai and chatgpt: Applications, challenges, and ai-human collaboration, 2023.

[Han *et al.*, 2021] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

[Hong and Curran, 2019] Joo-Wha Hong and Nathaniel Ming Curran. Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–16, 2019.

[Huang *et al.*, 2020] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J Cai. Ai song contest: Human-ai co-creation in songwriting. *arXiv preprint arXiv:2010.05388*, 2020.

[Joshi and Chakrabarty, 2021] Manish Joshi and Sangeeta Chakrabarty. An extensive review of computational dance automation techniques and applications. *Proceedings of the Royal Society A*, 477(2251):20210071, 2021.

[Lee *et al.*, 2023] Philseok Lee, Shea Fyffe, Mina Son, Zihao Jia, and Ziyu Yao. A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, 38(1):163–190, 2023.

[Liu *et al.*, 2020] Yuechang Liu, Dongbo Xie, Hankz Hankui Zhuo, and Liqian Lai. Plan2dance: Planning based choreographing from music. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13624–13625, 2020.

[McKight and Najab, 2010] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

[McKnight and Najab, 2010] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[Mirowski *et al.*, 2023] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34, 2023.

[Naveed *et al.*, 2023] Humza Naveed, Shi Khan, Asad Ullah afnd Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[Oliveira *et al.*, 2012] João Lobato Oliveira, Luis Paulo Reis, and Brígida Mónica Faria. Empiric evaluation of robot dancing framework based. *TELKOMNIKA*, 10(8):1701–8, 2012.

[OpenAI, 2023] OpenAI. Mastering temperature and top_p in chatgpt api. Technical report, 2023. Accessed January 2024.

[OpenAI, 2024] OpenAI. Six strategies for getting better results. Technical report, 2024. Accessed January 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Pignatiello *et al.*, 2020] Grant A Pignatiello, Richard J Martin, and Ronald L Hickman Jr. Decision fatigue: A conceptual analysis. *Journal of health psychology*, 25(1):123–135, 2020.

[Ramos *et al.*, 2015] Oscar E Ramos, Nicolas Mansard, Olivier Stasse, Christophe Benazeth, Sovannara Hak, and Layale Saab. Dancing humanoid robots: Systematic use of osid to compute dynamically consistent movements following a motion capture pattern. *IEEE Robotics & Automation Magazine*, 22(4):16–26, 2015.

[Sætra, 2023] Henrik Skaug Sætra. Generative ai: Here to stay, but for good? *Technology in Society*, 75:102372, 2023.

[Sagasti, 2019] Francisco Sagasti. Information technology and the arts: the evolution of computer choreography during the last half century. *Dance Chronicle*, 42(1):1–52, 2019.

[Saravia, 2022] Elvis Saravia. Prompt Engineering Guide. *https://github.com/dair-ai/Prompt-Engineering-Guide*, 12 2022.

[Scaringella *et al.*, 2006] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[Shinozaki *et al.*, 2007] Kuniya Shinozaki, Akitsugu Iwatani, and Ryohei Nakatsu. Concept and construction of a dance robot system. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts*, pages 161–164, 2007.

[Shinozaki *et al.*, 2008] Kuniya Shinozaki, Akitsugu Iwatani, and Ryohei Nakatsu. Construction and evaluation of a robot dance system. In *Entertainment Computing Symposium*, pages 83–94. Springer, 2008.

[Stiennon *et al.*, 2020] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[Sturm, 2012] Bob L Sturm. Two systems for automatic music genre recognition: What are they really recognizing? In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 69–74, 2012.

[Wang *et al.*, 2023a] Chi Wang, Susan Xueqing Liu, and Ahmed H Awadallah. Cost-effective hyperparameter optimization for large language model generation inference. *arXiv preprint arXiv:2303.04673*, 2023.

[Wang *et al.*, 2023b] Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112*, 2023.

[Wang, 2022] Zhigang Wang. Music choreography algorithm based on feature matching and fragment segmentation. *Computational Intelligence and Neuroscience*, 2022, 2022.

[Wei *et al.*, 2023] Chengwei Wei, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. An overview on language models: Recent developments and outlook. *arXiv preprint arXiv:2303.05759*, 2023.

[White *et al.*, 2023] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[Wu and Hu, 2023] Yangjian Wu and Gang Hu. Exploring prompt engineering with gpt language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, 2023.

[Zamfirescu-Pereira *et al.*, 2023] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[Zhou *et al.*, 2023] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.