PROJECT REPORT

Experiencing Co-Creativity: The Practice and Evaluation of Writing Variety Show Scripts with ChatGPT-3.5 by Industrial Professionals

Yifei Chen
Master of Arts in Computational Linguistics
University of Tuebingen
3 September, 2023

Key Words
Human-computer co-creativity, Text generation, Large language models

Course Title
Large Language Models: implications for linguistics, cognitive science and society

Word Count: 6980

Experiencing Co-Creativity:
The Practice and Evaluation of Writing Variety Show Scripts with ChatGPT-3.5 by Industrial Professionals

Abstract

This project is a first attempt on the application of large language models (ChatGPT-3.5) into co-creating variety show scripts, practising and evaluating by professionals. Currently, the development of language models has brought profound effects on the content creation industry but the realm of variety show production does not seem to be explored. Based on quantitative questionnaires and qualitative interviews, this project highlights the advantages and disadvantages of ChatGPT-3.5 in casting and script writing, envisioning a fine-tuned co-writing system in the future.

1 Introduction

The application of large language models in various forms of content creation such as visual arts and playwriting is a prominent research topic, particularly in the realm of story or narrative generation which has been extensively discussed (Rosa et al., 2020 & 2022; Yuan et al, 2022). The release of ChatGPT4 has taken this interest to another level, arising concerns among professionals in the creative industry including screenplay writers and book/magazine editors who become increasingly worried about the industrial future. A notable and significant event in this context is a series of strikes launched mainly by Hollywood writers and actors, including the *2023 SAG-AFTRA (Screen Actors Guild – American Federation of Television and Radio Artists) strike*. The primary reason behind it concerns a controversial contract that will guarantee the adoption of AI-generated screenplays and actors. However, there is a lack of structured review on the potential roles of those recent large language models in shaping the future of the industry, whether they pose a threat, serve as augmentative tools, or hold other possibilities. Notably, its role in the realm of variety shows which is a significant genre in the entertainment business has yet to be explored.

A most recent study is Mirowsky et al (2023), in which they invited 15 professionals in the theatre and film industry to co-write scripts using a newly developed system called Dramatron based on the *Chinchilla* large language model (Hoffmann et al 2022). Their study has demonstrated the possibilities and potentials of large language models in human-machine co-creativity on playwriting. This project is inspired by their work. However, in contrast to their work, this project looks at the variety shows instead of theatre plays, which is rarely researched under this topic. And additionally, while their study primarily focuses on evaluating the utility of Dramatron as a system specially designed for playwriting only, this project aims to closely examine the interactions between directors and large language models in the production process with script writing being the dominant part of it. Ideally, the outcomes of this project would serve as a valuable guideline for future developments on specialised variety show production systems based on large language models.

To elaborate, this project seeks to address the following questions:
1) To what extent can large language models be identified as a co-writer? Subsequently, what is the most accurate description for the potential role of current large language models during the production process?
2) How effective and creative can large language models be?
3) Recommendations for user experiences for the future developments.

2 Variety Show Production

2.1 Production Process

Variety shows represent a seamless blend of creative and realistic work, with its production process serving as a dual test of both physical and mental endurance. Production specifics could vary across diverse programmes such as talk show, travel show, and music programme. However, generally it involves three distinctive phases: 1) Pre-production, in this stage ideas and themes are conceptualised, scripts are written, and planning takes place for various elements such as set design and casting. 2) Production. For this stage filming according to the planned schedule is the main task. 3) Post-production, which is the polish process before the final content being delivered to the streaming media/television and presented to the audience. It includes editing and adding special effects and subtitles and more. These three stages collectively ensure the seamless show production.

Most studies in this field focus on the application of large language models in script writing which is a part of the production process, and overlook the potential for their implementation in other stages, like casting and idea generation during pre-production. This project will involve these two aspects to explore further possibilities of large language models in co-creativity.

2.2 Casting

The casting of a variety show is a crucial step in bringing the show to life. It involves the careful selection of individuals who will participate and engage in various tasks and challenges on the show. Participants will serve in various roles, including the host, main cast members, surprising guests, and experts from other fields may also be invited for educational purposes to input knowledge and background information. Members are expected to not only have a good sense of humour but, most importantly, an appealing personality that will evoke emotions and make the audience adore them. Producers and casting directors will be the main person responsible for this process. Participants are assessed and selected based mostly on their personalities, backgrounds, and their suitability to the theme and dynamics of the show. However, at times, under circumstances involving internal trade cooperation and opaque dealings, certain individuals can bypass these casting criteria and become participants directly.

2.3 Script Writing

Variety show scripts are left untouched while the play and drama script generation using (large) language models has been widely researched. This might be due to the unscripted and spontaneous nature of variety shows, which is considered as the defining characteristic that sets them apart from scripted television programmes and also plays. The interactions between participants, whether they are the main cast or guests, are placed in real-life situations thus not following pre-written lines or rehearsals. The genuine reactions and emotions provoked by given challenges or tasks are also unpredictable. These all lead to the fresh and dynamic viewing experiences ranging from humorous and heartwarming to dramatic and intense, which are exactly what make variety shows a compelling form of genre compared to other entertainments.

However, scripts are still required for shows for three main reasons: censorship, the need for a reference for cross-department collaboration and program scheduling, and certain segments still require careful planning and scripting such as choices of locations, design of programmes structure, design of challenges, in order to delicately guide the interactions and reactions of the participants and present a good show. In conclusion, the variety show script is a finely crafted work. It is intricately designed with deliberation and careful planning, while also leaving room for unplanned moments to unfold naturally. At the same time, it should appear to the audience as if there are no traces of a script.

The scriptwriting process is collaborative. Once the core concept (travel, talk, music, or other show themes) is established and the casting is nearly completed, the scriptwriting process begins. Taking outdoor shows as an example, after initial desk research and the selection of potential locations, the crew physically visits these sites to assess their feasibility and collect comprehensive information, including photos, audio recordings, and videos. Following this, they collectively brainstorm for segment designs, outlining content for each phase—usually divided into morning, noon, and evening time phases. Subsequently, directors draft the "structural script," encompassing not only content-related details such as program structure and segment design, but also information for censorship such as background notes and design rationale. Afterwards, the script writers tailor and create individual “role scripts” for each guest based on their personalities and assigned roles within the show.

The structure of a variety show script consists of a title and introduction to the show, subtitles for each episode, their introductions, and the content. The content of each episode generally contains the following basic elements: time, location, participants (especially the episode’s special guests), challenges, and rationale. A suggested template could be found in the project’s Github.

3 Methods

3.1 Large Language Models

This project employed the ChatGPT-3.5 developed by OpenAI to co-create the script. It is a large language model with a large-scale neural network containing 175 billion parameters (OpenAI 2023). These parameters are the core components that enable the model to understand and generate natural text across a wide range of natural language processing tasks. Its massive parameter count contributes to its ability to handle complex language tasks including text generation as applied in many studies, making it one of the most advanced language models available at present.

3.2 Interactive Writing Process

Tools Each participant is asked to use ChatGPT3.5 as an augmentative tool for their writing. They should use ChatGPT-3.5 only unless it is not sufficient enough to complete the script, and in which case the reasons namely the insufficiency of the ChatGPT-3.5 and the supplemented tools used should be noted. Due to sharing convenience and also a necessity to track changes that participants made to the original output produced by the model, the project adopts Google Docs as the writing platform. VPN is also applied as in mainland China participants do not have the access to Google and ChatGPT website. The VPN used is Cisco AnyConnect Secure Mobility Client.

Language The original language for writing, interviews, questionnaires and interactions with ChatGPT3.5 is Simplified Chinese. The following written materials are then translated to English for reference purpose only: the completed scripts, the transcripts of the interviews, and the questionnaires with responses.

Setting To ensure the coherence of topics across episodes, a basic framework was given. Specifically, participants were tasked with creating a script for one episode of a show based on a given scenario:

With the easing of lockdown and the relaxation of strict pandemic control policy, your company has decided to produce a new cultural travel show named "Let's Hang Out! (出来玩吧！)" to address market demands. In this show, the cast will embark on an exhilarating journey, *exploring different cities in China, with each episode focusing on one city. The participants include two main members who are already casted and two guest members featured differently each episode. Your job is to draft a structural script for the first/second/third/last episode in collaboration with ChatGPT-3.5. The script requirements and details are as follows:*
1) Each episode's maximum duration is limited to 90 minutes.
2) Cast two guest members.
3) Determine the location.
4) No budget restriction.

Introduction to the show: "Let's Hang Out!" is a captivating cultural travel reality show. In this program, four participants will embark on an exciting journey, travelling to four distinct Chinese cities, immersing themselves in the local culture and breathtaking landscapes. This endeavour is not just an exploration; it's an intimate encounter with the unique charm of each city. As the COVID-19 pandemic gradually recedes, we aim to engage the audience by bringing them into the participants' journeys, sharing in their joys, sorrows, and discoveries. Through the show, we encourage everyone to step out of their home and explore the world with friends and family.

Workflow Given that the project's primary focus is the performance of large language models as co-writers, there is no need for a time limit on the writing experiment. The professionals will engage in discussions with ChatGPT-3.5, refining their prompts if the initial conversation isn't progressing smoothly. And the process can be summarised as follows:

Stage	**Descriptions**	**Participants**
Tutoring	Tutorials for how to use ChatGPT-3.5	Project team
Preparing	Desk research: - Choose the city - and several candidate locations within the city Casting: two guest members Creating a title for the episode Designing the challenges	Professionals & ChatGPT-3.5
Writing	Writing scripts using the template	
Reviewing	Proofreading for grammar mistakes Adjusting the formatting Adding supplementary information Reviewing the rationale	Professionals, ChatGPT-3.5 & the consultant
Synthesising	Synthesising multiple individual episode scripts into a complete script	Project team & the consultant

Table 1. Project workflow

3.3 Participants

The participants are five professionals and ten variety show fanatics. They were all recruited through personal connections rather than open recruitment. Among the professionals, four will serve as co-writers while one acts as a consultant whose role is to assist in creating the template, overseeing script quality, and reviewing completed scripts.

All four professionals have at least one year of industry experience as directors, and to prevent potential bias they come from different companies. None of them have used ChatGPT before, although one professional (J.Li) has some experience in writing with AI. He has used the integrated AI in Notion (a software application and platform that provides a collaborative workspace for teams and individuals) which is developed based on OpenAI's ChatGPT to draft proposals, but not show scripts. As mentioned earlier, ChatGPT is not directly accessible in mainland China without VPN, which explains their lack of prior hands-on experience with the tool. However, they are at least aware of this transformative tool through videos shared on platforms such as TikTok and Bilibili (China's equivalent of YouTube).

The background and basic information of professionals are listed below with certain anonymity as they preferred:
1. *J. Li* (Male): a senior director with over a decade of experience and has founded his own production company.
2. *H. Cao* (Female): A senior member of a production company with over six years of experience.
3. *L. Zhou* (Female): Another senior member of a production company, with three years of experience.
4. *Y. Xia* (Female): A junior member of a production company, with one year of experience.
5. *J. Xie* (Female): the project consultant, who is a senior director position at a prominent streaming media company in China.

3.4 Evaluation Criteria

In this project, the professionals and the ChatGPT-3.5 were co-writing to generate a show script, making the creation process a collaborative one. This feature coheres with the notion of human-computer co-creativity as a collaborative creative process between a human and a computational agent (Kantosalo et al, 2016) and also a kind of creativity "distributed between humans and machines engaging in intense situations of symbolic interaction" (Assayag, 2021).

Human creativity is universally recognised as a valuable asset, and the creation of entirely novel and unexpected things is considered to be the zenith of human intelligence (Wiggins 2006). And creativity holds equal importance for large language models as for humans (Boden, 2003). In alignment with human creativity, which is defined as "the ability to come up with ideas or artefacts that are new, surprising, and valuable" (Boden, 1994), any task that, if executed by a human, would be acknowledged as requiring this skill could also be considered creative for AI (Wiggins, 2006), so as well as for large language models within our context.

Most attempts to evaluate the computer agent's creativity, from pioneers like Boden (1977, 1990), have focused on seeking creativity generated by or fostered via algorithmic means. Boden addressed the challenges of AI's creative behaviour and introduced a descriptive hierarchy of creativity. Building upon this, Wiggins (2006) proposed a framework for the formulation of creativity, aiming to develop a model that allows for a detailed comparison and thus a better understanding of systems that exhibit behaviours regarded as "creative" in humans. Nevertheless, as this project does not seek for algorithmic measures, how could creativity be evaluated?

The presence of value (or usefulness) and novelty, have been identified by many researchers in this field as two fundamental aspects defining a creative process (Boden, 2003). An automated generative system has the capacity to extensively explore various new combinations of elements, which can often lead to outcomes or creations that lack significant interest. Consequently, computational creativity necessitates not only the production of novel creations, but also ones that hold value. While other facets of creativity have been deliberated and put forth (surprise, for example (Macedo & Cardoso, 2001)), the prevailing commonalities embraced by most theories in computational creativity are the elements of novelty and value. Therefore, for this project creativity will be evaluated by these two dimensions.

As co-creativity is a collaborative process, the relationships between human beings and large language models are worth exploring. In this regard, Lubart's (2005) has proposed a categorisation of creative computational collaborators into four distinct roles that effectively exemplifies the notion of computational agents as tools: nanny, pen-pal, coach, and colleague. When applied to the context of large language models (LLM) in this project, these roles could be adjusted respectively as follows:
1) Intern/Secretariat (Similar to nanny): LLM assists the professional's tasks and time allocation for creative activities, while also handling routine responsibilities like saving and presenting information;
2) Consultant (Similar to coach): LLM advises professionals of techniques that promote creativity to stimulate their creative process.
3) Colleague: LLM exhibits creativity on its own, or it can contribute fresh ideas through a dialogue with professionals.

Eventually, incorporated with the evaluation criteria from the inspiring study by Mirovsky (2023), the evaluation design (the questionnaires and interview outlines) could be summarised in Table 2 below, and all the quantitative questions are 5-point Likert scales from strongly disagree to strongly agree:

Value/Usefulness
Quantitative
Q1: Correctness *Q2: Helpfulness* *Q3: Effectiveness*
Qualitative outlines
Q1: How do you collaborate with ChatGPT-3.5? Please guide me through the entire process, explaining each stage. *Q2: What tasks did ChatGPT-3.5 excel at? Among its accomplishments, which one stood out the most? What specific application do you find (most) remarkable?* *Q3: What areas could ChatGPT-3.5 enhance? And among these, what aspect left you the most disappointed or frustrated?* *Q4: When did you encounter challenges while writing with ChatGPT-3.5? Have you encountered situations where you needed to ask multiple questions to obtain a satisfactory response? If so, was it due to the complexity of the question or some other reason?* *Q5: Are there any concerns related to stereotypes or biases in the responses generated by ChatGPT-3.5?* *Q6: Have there been instances of incorrect information being provided?* *Q7: Are there any concerns regarding unusual phrasing, grammar mistakes in the responses?* *Q8: Have there been cases where the coherence or relevance of the content was lacking?* *Q9: Anything else about its usefulness? (Open discussion)*
Novelty
Quantitative
Q1: Uniqueness

Q1: Inspiring
Qualitative outlines
Q1: Has ChatGPT-3.5 produced any innovative ideas? *Q2: Has it been a source of inspiration for you to generate new and innovative concepts?* *Q3: Have you identified any instances of plagiarism in the content generated by ChatGPT-3.5?* *Q4: Anything else about its creativity? (Open discussion)*
Relationships
Quantitative
Q1: Collaborativeness (4 roles) *Q2: Enjoyment* *Q3: Satisfaction* *Q4: Pride* *Q5: Authorship*
Qualitative outlines
Q1: How would you describe the relationships between you and ChatGPT-3.5? *Q2: While co-writing with ChatGPT3.5 (which likely handles a significant portion of the detailed work), do you still perceive the output as your own creation? (Authorship)* *Q3: Throughout the entire collaboration process, what are your feelings and impressions about ChatGPT-3.5? From the beginning until now, have there been any changes in the overall experience?* *Q4: Do you have any recommendations for the user design that future developers could implement when designing a writing system for show scripts?* *Q5: Open discussion on relationships.*
Output Quality
Quantitative
Q1: Humanoid *Q2: Readiness* *Q3: Attractiveness* *Q4: Fun* *Q5: Uniqueness*
Open Section for further comments.

Table 2. Evaluation sheet

4 Results

Usefulness

Table 3	Average	Li	Cao	Zhou	Xia
Correctness	2.75	3	3	2	3
Helpfulness	2.75	2	2	3	4
Effectiveness	2.00	1	2	2	3
In total	2.50	2.00	2.33	2.33	3.33

Table 3 presents the results for the quantitative scales. The low scores suggest that the professionals did not find the tool to be particularly useful. All of them followed the suggested topic path: location – casting – challenges – modification. One unique case is Zhou, who opted not to seek assistance from ChatGPT-3.5 for city selection. Instead, she chose Chengdu directly due to her local familiarity, out of curiosity to explore if the response would be acceptable to her and how much it would deviate from her knowledge.

All of them acknowledge that ChatGPT-3.5 is helpful for desk research. “*GPT did a great job when trying to sort out information at the beginning,*” said Zhou. Xia also expressed something similar: “*[...] helps me to save a lot of time when searching information as it could understand my needs better than traditional search engines.*”

Another frequently mentioned advantage of ChatGPT-3.5 is its usefulness in generating content that doesn't require much intelligence but is time-consuming, such as stringing words together, dialogues, official introductions to the show/location, official jargon, and so on. However, they did mention that when it comes to writing certain political official languages, ChatGPT-3.5 was not competent. This is a matter that has regional and cultural specificity thus could be tolerated. Upon my observation, this is likely why junior professionals assigned higher scores for helpfulness. Juniors dedicate more time to these tedious tasks compared to the senior counterparts.

All professionals acknowledge that ChatGPT-3.5 is fully capable of handling such writing tasks or at least requiring much less revisions compared to other types of assignments. For instance, when designing dialogues, ChatGPT-3.5 defaultly assumes the presence of a host, which is a common scenario. But once prompted about the absence, it swiftly adjusts its response and provides smooth conversations among the guests with all the key messages still being covered. Moreover, ChatGPT-3.5 showcases its proficiency in handling specific requirements, such as determining the sequence of cast members' appearances, incorporating particular content (like location introductions) into dialogues, and more. "*I find this quite tricky - having large language models that are based on natural language generate natural language*." Xia remarked. Additionally, professionals were pleasantly surprised that ChatGPT-3.5 performed well in using figurative language. If requested, ChatGPT-3.5 could provide a beautifully written description of a location or cultural insight, enriched with figurative languages such as metaphors and parallelism. “*A bit formulaic. I mean cliché, but it gets the job done. After all, it (the script) is not meant for the audience, but for the decision makers and censors*.” Said Xia. And all professionals agreed that there were mere grammar mistakes.

At times during the conversations, ChaGPT-3.5 could exhibit a certain inclination to think like a director. For example, “*when asked to ‘elaborate on the value of Guiyang city from the perspective of show production,’ GPT could even understand that my purpose is to evaluate if Guiyang city is a suitable filming location and answer this question by suggesting various challenges or themes we could consider based on the uniqueness of the city.*” Said Li. Another evidence about ChatGPT-3.5’s potential to think like a director is that sometimes he would give production advice even without asking. Cao and Li both mentioned that ChatGPT-3.5 would attache some production tips when being asked other irrelevant questions: “*GPT said something like ‘it’s best to get in touch with the management personnel before visiting the location.*’”

While professionals all agreed on the general helpfulness of ChatGPT-3.5 in desk research, they also named severe shortcomings. Firstly, the output generated by ChatGPT-3.5 is limited to textual format, which is inadequate for variety show scripts. Unlike screenplay or play scripts, which are narrative-driven, variety show scripts serve more as production guidelines for which require visual elements like images and videos are vital, as they assist directors in assessing whether the location or culture is suitable for filming such as its visual appeal and attractions to the audience. For instance, if a location possesses intriguing traditional cultural elements like traditional dance, music, or handicrafts, mere text is insufficient for comprehensive assessment and supplementary video or image material are crucial. Professionals would have to reach out to external resources such as Bilibili and Weibo (China's equivalent of Twitter) to obtain. As Li candidly remarked, "Frankly, this kind of desk research is a crucial aspect of pre-production, as it's essential to have a reasonably deep understanding of local cultural content before embarking on on-site exploration." Consequently, it is evident that ChatGPT-3.5's current exclusive provision of textual output falls short of fulfilling these demands.

The conception and creative contents in the show are significantly rooted in practical fieldwork, which means the writing process involves not only extensive desk research but also (and arguably more of) rigorous inspections in the field. For instance, challenges might need to be tested and modified several times to prove their feasibility and logic. Moreover, selecting suitable filming locations for outdoor variety shows often entails comprehensive on-site inspections. Factors such as the appearance of the site, its suitability for shooting, permissions for filming, cooperation from local authorities or relevant organisations, the cost and fees, and transportation convenience all need to be taken into account. All of these are way beyond ChatGPT-3.5's capability.

Casting is also argued to be beyond ChatGPT-3.5's capabilities, as it requires timely and insider information. This includes traffic data on social networks and even internal transactions, as well as other confidential details. Moreover, the candidates ChatGPT-3.5 provided were mostly well-known people, with little possibilities on those who are not that famous. Not to mention that the information it provides is often incorrect. For instance, when asked for notable works of an artist, all those listed were incorrect. Additionally, it lacks information on candidates who have either faced legal issues or have tarnished reputations (not necessarily illegal but morally questionable), and hence are not recommended to appear in front of the public eye anymore.

Several additional challenges were also identified. Given that most professionals lacked prior experience with large language models, offering appropriate prompts or instructions became challenging. It typically took them 3-5 attempts to receive an acceptable response. In many cases, a considerable amount of information needed to be explained or even taught to ChatGPT-3.5. However, such basic knowledge is often taken for granted for practitioners "*[...]unless it's an intern, and I mean like a super beginner level intern*." quoted Cao. For instance, it is a basic consensus that the distance between filming locations within one episode should not be too far apart and this distance is presumed to be within an one-hour drive, as anything farther could lead to significant time waste during transitions. But ChatGPT-3.5 is not aware of this as it lacks the industrial consensus, thus leading to low efficiency.

Speaking of industrial consensus, ChatGPT-3.5 seems not to possess a template for a variety show script. Although it understands what a variety show script is and what components it should include, it lacks the ability to compose each part like a human and form a complete script. Furthermore, it is also unable to truly understand each component and not to mention generate the according content. For instance, while it knows that a variety show script includes the program flow, "*if you ask it to write the rundown/agenda, it either ends up composing dialogues or producing lyrical prose, using too much adjectives and modifiers*." Said Xia.

Last but not least, a frequently mentioned disadvantage is ChatGPT-3.5's limitation in maintaining a broad picture, resulting in coherence and logical consistency issues. The collaboration process with ChatGPT-3.5 is conducted part by part, which means it may not retain the overall conceptual design. For instance, in Zhou's script, she conceptualised the episodes as a role-playing game where guests became players who need to collect three valuable human artefacts to defeat cyber invaders coming from the future world and protect our planet. However, during the whole writing process, ChatGPT-3.5 does not inherently retain this overarching concept unless being explicitly emphasised and reminded by Zhou. "*It's like teaching a really slow intern who just can't get it, three-minute memory, might as well do it myself and that's much faster*." Cao complained.

Incorrect information is not a rare case. Some of them are relatively normal. For instance, as the latest data of ChatGPT-3.5 was collected by September 2021, certain recent updates on contact numbers, website links and guest background information might be inaccurate. However, some misleads are quite absurd. Despite having committed legal violations prior to 2021, some celebrities were still recommended by ChatGPT. And ChatGPT's mistakes regarding some very basic information such as the birthplaces of most potential participants are astonishing.

Another issue regarding this is that information provided by ChatGPT-3.5 does not have convincing resources. When asked what the references were, ChatGPT-3.5 can't name or list them clearly. So professionals need to spend tremendous time checking it. "*I really doubt if GPT could really save our time, simple information that used to be written by humans rarely went wrong, but now we have to spend extra time double-checking it. And for complex information, we needed to check it anyway*." Zhou sighed.

Novelty

Table 4	Average	Li	Cao	Zhou	Xia
Uniqueness	1.75	1	2	2	2
Inspiring	2.50	2	2	3	3
In total	2.13	1.50	2.00	2.50	2.50

"*I don't think GPT really has the ability to innovate.*" Almost all professionals expressed this opinion regretfully, as the challenges it designed are mostly quite common. However, since ChatGPT-3.5 is proficient in generating a large volume of outputs, which means it has the ability to consistently deliver over 10 potential challenges as required and cover a diverse array of categories. Therefore, it can be utilised with reservation as "a pool of inspirations to some extent" (Zhou). "Reservation" as this pool is not authentically a repository of variety show challenges. While ChatGPT-3.5 does indeed collect data about some challenges from variety shows, they often lack details or uniqueness. For example, "hide and seek" is a universally known challenge that not is not unique for variety shows. Subsequently, this pool was not specially tailored for variety shows and lacks the nuanced fine-tunes, making its application in the creative process comparatively weak.

Furthermore, attempting to get ChatGPT-3.5 into providing a comprehensive conceptual design (such as the role-playing game setting mentioned earlier) is nearly an impossible task. When presented with an example, ChatGPT-3.5 tends to build its design solely from that example, failing to grasp the intent of breaking conventions and crafting settings that are characterised by surrealism or novelty.

The issue of plagiarism is not criticised. The main reason is that challenges could be pretty similar across different shows as most of them are based on universal knowledge or games.

In sum, ChatGPT-3.5 does not seem to demonstrate a satisfactory creativity with negative reviews outweighed positive ones. The reasons behind this are in accord with previous literatures,

Relationships/User Experience

Table 5	Average	Li	Cao	Zhou	Xia
Collaborativeness	/	Secretariat/Intern	Secretariat/Intern	Secretariat/Intern	Consultant
Enjoyment	2.75	2	3	3	3
Satisfaction	3.00	2	3	3	4
Pride	2.25	1	2	3	3
Authorship	3.75	4	4	4	3
In total	2.94	2.25	3.00	3.25	3.25

Li characterised ChatGPT-3.5's role as the secretariat/intern: "*[...] probably the worst one I ever hired*," he joked, "*definitely won't pay for the salary*." He mainly employed the GPT-3.5 for generating routine content, such as background information and dialogues, as well as for conducting early desk research. This was also his principal purpose for using Notion AI. "*The only reason I would hire it (GPT) is that I need someone who knows about the English world, (as) I can't speak the language*," Li further emphasised. Zhou echoed this comment, explaining that traditional search engines perform poorly when attempting to access English information, both in terms of quality and quantity. And due to a lack of language proficiency, she struggles to effectively filter through the results. "*I could use an intern for that, absolutel*y." she added. These three professionals including Cao shared the same opinion regarding ChatGPT-3.5's role as an intern. This perspective stems primarily from ChatGPT-3.5’s limited competences of industry-specific knowledge and the interactions being more similar to superiors delivering instructions to subordinates, rather than fostering inspiring discussions or brainstorms among colleagues. Only Xia considered ChatGPT-3.5 to be a consultant, as it helped her with the reluctant basic desk research and provided diverse information in languages other than Chinese to offer her inspiration. This process is similar to having a foreign consultant. However, she also added that it must be “*a consultant with limited Chinese comprehension skills*” who required repeated consultations and simplified questions. Overall, professionals reached the agreement that the interactions leaned more toward question-and-answer rather than a back-and-forth discussion.

Regarding their emotional experience during co-creation with ChatGPT-3.5, on average, they did not feel much enjoyment, satisfaction, or pride. The reasons behind are then illuminated in the interviews. Except for Li, the others generally found the process enjoyable and satisfying for several reasons. At the beginning, the novelty and curiosity about the new tool contributed to a sense of enjoyment. However, as they delved deeper into the process, the novelty wore off, surprises dwindled, and they realised the extensive amount of manual work required such as correcting errors in formatting and wrong information given. Consequently, the overall enjoyment score was not high. Furthermore, while the quality of the final script was acceptable, it was only moderately satisfactory, and for professionals this level of quality certainly could not reach a level of pride. Li's score was even lower due to his emphasis on efficiency. The time and effort invested in the co-creation process with ChatGPT-3.5 did not match the output quality obtained. Before assigning scores, I had already informed them that the process of copying and pasting from ChatGPT-3.5's output into Google Docs was not within the scope of evaluation, as it's an aspect that can be quickly improved if there is a writing system. Nonetheless, they highlighted its significance, as the copy-paste process can be patience-consuming.

Since the content generated by ChatGPT-3.5 is somewhat templated and fragmented, professionals took on the task of interconnecting and refining the details. "It's like having a lot of raw materials (very likely not quite reliable), and then I craft the final product," described Cao. Consequently, all participants awarded higher scores for authorship, viewing the output as essentially their own creation.

Quality

Table 6	Average	Highest	Lowest
Humanoid	5.60	8	4
Readiness	7.70	9	6
Attractiveness	4.60	6	4
Fun	4.60	6	4
Uniqueness	2.70	4	2
In total	5.04	/	/

Overall, the results show that in general fanatics found the quality of the scripts acceptable. The AI's influence was not overly pronounced; the content was coherent and highly readable. However, the plots and designs were considered conventional and lacking in originality and novelty. If aired, the scripts wouldn't hold much appeal unless the guests were personally interesting to them, as one of the interviewees stated. This conclusion aligns with the feedback above by professionals regarding the scripts and the collaboration process.

5 Future Implementation

As highlighted in the introduction, this project envisions the integration of large language models into variety show production systems, at least for the script writing process. All participants in this project share the belief that such a system will become a reality, given the prevailing trend of AI application across various industries. Professionals expect that the system excel in the following functions:
1) Capability to generate introductory and other routine paragraphs.
2) Ability to provide accurate information without grammatical errors.
3) Aptitude to maintain script coherence in line with the provided concept.
4) Proficiency in generating dialogues as required.
5) Ability to automatically produce output in the professional template format.
6) Inspiring the creation of challenges through a vast dataset of challenges covers different kinds of variety shows worldwide.

6 Conclusion

This project is inspired by Mirowsky et al (2023) and aims to evaluate the feasibility, effectiveness, and creativity of utilising ChatGPT-3.5 to co-write variety show scripts with professionals from junior to senior levels. And this project is arguably the first attempt to evaluate the application of large language models into variety show production. Through the engagement of the collaborative process and subsequent interviews, I have gained valuable insights into the potential of this collaboration. Throughout the research, a range of strengths and limitations of ChatGPT-3.5 in the realm of variety show content creation are revealed.

Firstly, ChatGPT-3.5 demonstrated significant advantages in supporting early desk research and generating routine content. Professionals acknowledged its valuable assistance in tasks such as information collecting and dialogue generation, which contributed to workload reduction and time savings, thus having the potential to free practitioners from routine work. However, its ability of creativity and innovation was unfortunately limited as either show fanatics or professionals found the content quite innovative.

Furthermore, from the feedback of industry experts, the challenges posed by accuracy and novelty in ChatGPT-3.5's output are also observed. While it can produce substantial content, the generated material might contain factual inaccuracies, formatting issues, and more. Moreover, due to its mechanical and fragmented output, professionals often needed to invest additional effort in reorganising and refining the content.

Despite these limitations, the potential of ChatGPT-3.5 in providing inspiration and guidance is also identified. It has the potential to serve as a creative starting point, offering diverse options and directions for professionals to explore. But in order to make it a reality, finetune and further machine training are required, which could be a future direction.

In summary, the findings of this project call for further exploration of AI applications in the realm of variety show production. Although ChatGPT-3.5 exhibits limitations in terms of innovation and human-like creativity, large language models still hold significant potential as an auxiliary tool to enhance productivity and broaden creative horizons, especially after the release of ChatGPT-4. Future research endeavours could delve into refining collaborative models, enhancing large language model's creativity and efficiency in scriptwriting across diverse genres of variety shows, not only restricted to outdoor shows as in this project.

7 Ethical Consideration and Limitations

Before the professionals and fanatics engaged in the study, informed consent was obtained from each of them. They were provided with comprehensive information about the research objectives, procedures, potential risks, benefits, and their rights as participants. Participants had the option to withdraw at any point without consequence. The confidentiality of participants was important as well. All identities of fanatics were anonymised, and any personal or identifying information was kept confidential. Participation in the study was entirely voluntary. Professionals were under no obligation to take part and additionally, their decisions to participate or withdraw did not impact their professional relationships or opportunities.

This project presents several limitations. Firstly, the language employed for collaboration with ChatGPT is Mandarin, with an accuracy rate of 80.1%. As ChatGPT's performance varies across languages (OpenAI, 2023), this limitation might restrict the reliability of the findings. The outcomes could potentially be more robust if using languages with better performance, such as English, which possesses an 85.5% accuracy rate. Another limitation pertains to the sample size. Participants in this project were invited by personal connections rather than being chosen randomly. This non-random selection process introduces the possibility of information cocoons. Furthermore, due to restricted industry connections, no supplementary list was available for the selection of participants based on specific criteria, thus might result in certain bias. Thirdly, this project is founded on GPT-3.5 while OpenAI has already released the newest version - GPT-4. It is plausible that GPT-4 may exhibit better performance, potentially yielding new insights that are different from those presented here.

In light of these limitations, the findings of this project should be interpreted within the context of these constraints. Future researchers could consider addressing these limitations to provide a more comprehensive understanding of co-creativity in the entertainment industry.

Reference

Assayag, G. (2021). *Human-Machine Co-Creativity*. In Artisticiel / Cyber-Improvisations (pp.ffhal-03542917ff, 2021. Ffhal-03542986f). Phonofaune.

Boden, M. (1977). *Artificial Intelligence and Natural Man*. Harvester Press.

Boden, M. (1990). *The creative mind*. Abacus.

Boden, M. (1994). *Modelling creativity: reply to reviewers*. Journal of Artificial Intelligence, 79, 161-182.

Boden, M. A. (2003). *The creative mind: Myths and mechanisms*. Routledge.

Gervás, P., Díaz-Agudo, B., & Peinado, F. (2005). *Story Plot Generation Based on CBR*. Knowledge-Based Systems, 18(4), 235-242.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., & Huang, Z. (2022). *Training Compute-Optimal Large Language Models*. arXiv preprint arXiv:2203.15556.

Kantosalo, A., & Toivonen, H. (2016). *Modes for Creative Human-Computer Collaboration: Alternating and Task-Divided Co-Creativity*. In International Conference on Innovative Computing and Cloud Computing.

Lubart, T. (2005). *How can computers be partners in the creative process: classification and commentary on the special issue*. International Journal of Human-Computer Studies, 63(4), 365-369.

Mirowski, P., Yan, X., Sifre, L., Razavi, A., Vinyals, O., Viola, F., & Klimov, O. (2022). *Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals.* ACM Transactions on Interactive Intelligent Systems (TiiS), 12(2), 1-27. https://doi.org/10.1145/3544548.3581225.

Macedo, L., & Cardoso, A. (2001). *Modeling forms of surprise in an artificial agent*. Proceedings of the 23rd Annual Conference of the Cognitive Science Society.

OpenAI. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/arxiv.2303.08774.

Rosa, R., Honko, H., Kanerva, J., & Luukka, P. (2022). *GPT2-based Human-in-the-loop Theatre Play Script Generation*. In Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22) (pp. 841-852). Association for Computing Machinery. https://doi.org/10.1145/3490099.3511105.

Rosa, R., Honko, H., Kanerva, J., & Luukka, P. (2020). *THEaiTRE: Artificial Intelligence to Write a Theatre Play*. arXiv preprint arXiv:2006.14668.

Wiggins, G. A. (2006). *A preliminary framework for description, analysis and comparison of creative systems*. Knowledge-Based Systems, 19(7), 449-458.

Yuan, A., Yang, Z., Gu, Q., & Yeo, C. K. (2022). *Wordcraft: Story Writing With Large Language Models*. In Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22) (pp. 841-852). Association for Computing Machinery. https://doi.org/10.1145/3490099.3511105.

Appendix

All the interview audio recordings, transcripts, and evaluation sheets could be found on the project's Github page: https://github.com/devychen/llmcoursework/tree/main.