


How good is ChatGPT? An exploratory study on ChatGPT's performance in engineering design tasks and subjective decision-making

Wanyu Xu , Maulik Chhabilkumar Kotecha and Daniel A. McAdams

Texas A&M University, United States of America

 wanyu.xu@tamu.edu

Abstract

This study explores how large language models like ChatGPT comprehend language and assess information. Through two experiments, we compare ChatGPT's performance with humans', addressing two key questions: 1) How does ChatGPT compare with human raters in evaluating judgment-based tasks like speculative technology realization? 2) How well does ChatGPT extract technical knowledge from non-technical content, such as mining speculative technologies from text, compared to humans? Results suggest ChatGPT's promise in knowledge extraction but also reveal a disparity with humans in decision-making.

Keywords: artificial intelligence (AI), design knowledge, design research, design theory

1. Introduction

Advancements in artificial intelligence, particularly in the field of natural language processing (NLP), have revolutionized the way machines interact with and understand human language. One significant milestone in this field is the development of large language models (LLMs) (Tom *et al.* 2020; Hoffmann *et al.* 2022; Zhang *et al.* 2022). These models are based on deep learning architectures, specifically variants of the transformer architecture, which have demonstrated exceptional proficiency in processing and generating text (Ray 2023). Notably, ChatGPT (OpenAI 2022), a chatbot powered by generative pre-trained transformer (GPT) models that are trained on an extensive corpus of diverse text from the internet, stands out for its ability to generate coherent and contextually appropriate responses to a wide range of prompts (Koubaa 2023). These pretrained generative language models exhibit remarkable capabilities in zero-shot, one-shot, and few-shot learning scenarios, contributing to the increasing popularity of GPT models among researchers exploring domain-specific applications.

Currently, most research on the application of GPT in engineering design primarily focuses on its use in the early stage of the design process. Existing studies have shown promising results in this area, including concept generation in collaboration with human designers through GPT-powered chatbots or by GPT models alone (Wang *et al.* 2023; Zhu and Luo 2023). Before further integrating generative large language models into our design process, it is crucial to understand how well these models, such as ChatGPT, perform, compared to humans, in understanding and responding to textual information in engineering design tasks. However, before further integrating generative large language models into our design team, evaluating their performance, particularly in comparison to humans, in comprehending and responding to textual information within engineering design tasks is imperative. Addressing this question is essential for understanding the disparity between large language models and human designers, facilitating collaboration between humans and artificial intelligence in engineering design processes.

We are particularly interested in evaluating the effectiveness of models like ChatGPT in extracting technical content from non-technical sources and assessing judgment-based tasks. This idea is inspired by our activities in manually collecting and verifying speculative technologies from science fiction narratives for design ideation and decision-making research. The manual process of technology realization evaluation was labour intensive. To assess ChatGPT's performance relative to humans, we assigned ChatGPT the same task that our human team members completed in that science-fiction related study (Xu *et al.* 2021). Through this exploratory study, we aim to answer the following research questions:

1. *How does ChatGPT compare with human raters in evaluating judgment-based tasks such as speculative technology realization evaluation?*
2. *How good is ChatGPT at knowledge discovery of a technical nature from non-technical content such as mining speculative technologies from a given text as compared to humans?*

The rest of this paper is organized as follows: Section 2 provides a brief review of recent studies on generative pretrained LLMs in the early product development stage and prior work on machine evaluation of subjective evaluation tasks and decision-making. Section 3 presents the details of the study to answer the research questions. Section 4 consists of the results and observations, followed by a discussion on the implications of the observations. Finally, in Section 5, we present our conclusions and future directions.

2. Literature review

2.1. GPTs in the early product development stage

Generative Pretrained Transformers (GPT), one of the most prominent models in natural language processing, are well known for their capability to understand and generate content. Trained on vast amount of data, GPT models can perform a broad range of tasks with domain adaptability and versatility. In recent years, there has been a growing interest in the application of GPTs during the early stages of product development. Siddharth *et al.*, in their review of natural language processing techniques in the context of design research, highlight the future opportunities of GPT models in engineering design, including concept ideation, concept search, concept association, and concept retrieval (Siddharth *et al.* 2021).

Zhu and Luo (2023) pioneered in experimenting with GPT-2 and GPT-3 to explore GPT's capability in data-driven concept generation for early-stage design. Furthermore, their work in bio-inspired design demonstrated how fine-tuned GPTs could generate feasible and novel concepts, helping engineering teams in concept selection for further development. Building upon this foundation, Zhu *et al.* expanded the application of GPTs in the early design stage by finetuning GPTs for both concept generation and concept evaluation in bio-inspired design. Besides demonstrating GPT's performance in generating feasible and novel bio-inspired design concept, this work showed that GPTs could be customized to retrieve the knowledge and reasoning for biologically inspired design process, aiding designers with limited expertise (Q. Zhu *et al.* 2023). Additionally, researchers such as Ma *et al.* have explored GPT-3's capabilities in generating design solutions and compared them with crowdsourced solutions in terms of feasibility, usefulness, and novelty (Ma *et al.* 2023). Their findings indicated that while GPT-3 generated solutions were more feasible and useful, they lacked novelty compared to crowdsourced solutions. Furthermore, evaluation into the influence of prompts on design outcomes revealed that the few-shot prompt yielded set of solutions most similar to those from crowdsourced workers according to their computational Evaluation metric.

Addressing the challenges of transparency and controllability in GPTs, Wang *et al.* introduced a task-decomposed AI-aided approach for generative conceptual design, inspired by the Function-Behavior-Structure (FBS) model (Wang *et al.* 2023). Their methodology decomposes design tasks into functional, behavioural, and structural reasoning sub-tasks, utilizing prompt templates and specification signifiers to guide LLMs for generating reasonable output. Experimental findings demonstrated that their approach led to more reasonable and creative design ideations compared to a baseline with a free-form

prompt. Additionally, this proposed approach improved performance among designers of different experience levels, particularly benefiting novice designers.

Research on GPT applications in engineering design is still in its early stages, requiring further exploration in areas such as Human-AI comparison studies, alignment research, and the development of frameworks, guidelines, and collaborative strategies between designers and the generative AI.

2.2. Machine evaluation of subjective evaluation tasks and decision-making

We reviewed the literature where LLM-based tools are used to evaluate subjective information and compare its performance against human evaluators. [Zhu et al. \(2023\)](#) performed an annotation task using ChatGPT and compared it with human annotation for various types of datasets such as stance detection, sentiment analysis, hate speech, and bot detection and reported an average accuracy of 0.609. Ji et al. conducted multiple tests to assess the ability of ChatGPT to rank content and concluded that the ranking preferences of ChatGPT are consistent with humans ([Ji et al. 2023](#)).

Overall, the advancements in artificial intelligence, specifically in the domain of natural language processing, have paved the way for tools like ChatGPT to play a crucial role in the engineering design. As design researchers, we are interested in understanding how large language models (LLMs) like ChatGPT and humans understand a language and work together to achieve common goals. Our research explores the intersection between artificial intelligence and human evaluation.

3. Study design

A broader question that we explore through our research is how good language models are as part of a design research team and as a decision-maker. To that end, we begin with the following two research questions:

1. *How does ChatGPT compare with human raters in evaluating judgment-based tasks such as speculative technology realization evaluation?*
2. *How good is ChatGPT at knowledge discovery of a technical nature from non-technical content such as mining speculative technologies from a given text as compared to humans?*

To answer these questions, we designed two experiments based on a previous study completed in 2021 ([Xu et al. 2021](#)). Researchers have explored the potential of leveraging science fiction narratives to enhance designers' ideation processes ([Sterling 2005](#); [Bleecker 2009](#); [Dunne and Raby 2013](#); [Callaghan 2015](#); [Kotecha et al. 2021](#)). However, to fully incorporate science fiction into design, we must deal with the time-consuming and labour-intensive task of collecting and verifying speculative technologies from science fiction narratives. In the previous study, we organized a large set of speculative technologies as introduced in speculative fiction from an online repository called *Technovelgy* ([Christensen 2019](#)). There are 3,094 technologies in the *Technovelgy* dataset in 32 different categories and each entry in the dataset contains the name of the speculated technology, a short description of the technology, and the corresponding excerpt where the technology appears. We then employed five human raters to determine if the technologies were realized so that each technology was reviewed by three different human raters. Further details about this study can be found in a previous paper ([Xu et al. 2021](#)).

In the current study, we asked ChatGPT to complete the same task as what our human team members performed in ([Xu et al. 2021](#)) and compared the outcomes against each other. The study was conducted under the assumption that ChatGPT had similar accessible resources as our human raters in early 2021, given that current GPT models were all trained on data up to September 2021 ([OpenAI 2022](#)).

Our experimental outcomes were derived from submitting a large volume of queries to GPT-3.5-Turbo-1106 model via the ChatGPT API. For this exploratory study, we adopted the default parameter settings for top P, frequency penalty, and presence penalty, which were 1, 0, and 0, respectively. We changed the temperature to 0 to reduce randomness and increased the maximum token length to 1024 to accommodate the requirements of our task.

Each query contains two types of information, as shown in Figure 1 and Figure 2: system information and user input. In each experiment, every query shares the same system information, which specifies the persona of the model, explains the task, and provides an example of the assigned task. The user input, on the other hand, contains entry-specific details. For task one, it includes an excerpt from science

fiction, whereas for task two, it includes the name and a concise description of a speculative technology, along with a corresponding excerpt from science fiction where the technology is mentioned.

3.1. Experiment 1: technology identification

Task: ChatGPT was given the following prompt and an excerpt from science fiction as shown in Figure 1. Each excerpt contains at least one speculated technology from the technology dataset. ChatGPT was asked to identify whether there was a technology in this excerpt. If so, ChatGPT needs to name the technology and write a brief description of the technology. Responses were separated into technologies and descriptions and the running time for each query was recorded.

Measure: For those successfully identified entries, we compared the semantic similarity between every ChatGPT-generated technology description and the corresponding human-generated technology description from the *Technovelgy* dataset. We chose to use the Bidirectional Encoder Representations from Transformers (BERT) model, which is known for its ability to capture rich contextual information and semantic representations of language, to convert descriptions to embeddings and then calculate the cosine similarity between the embeddings. We also randomly selected 20 entries from the dataset and asked human raters to generate technology descriptions. The semantic similarity between the original description from the dataset and the newly generated ones were used as a benchmark for ChatGPT's performance.

System information:

You will be presented with some excerpts from speculative fiction and your job is to identify the speculated technology from the text and provide a brief description of the technology (2 sentences at most). Provide your answer in the same format as the following example. If you cannot find any technology, then say "I don't know".

Example:

User: "One of the most successful of these various contrivances, and the one, indeed, in which I was most deeply interested, was a small machine very much resembling in appearance the tube, with a mouth-piece at one end and an ear-piece at the other, frequently used by deaf persons, but very different in its construction and action. In the ordinary instrument the words spoken "This translation was accomplished by means of certain delicate machinery contained in the end of the mouth-piece, which was longer and larger than that of the ordinary ear-tube, but the

...

... to see how it would work with some other person than myself at the mouth-piece. In the course of its construction I had frequently tried the machine by putting the ear-piece into my ear and speaking into the mouth-piece such scraps of foreign languages as I was able to command. These experiments were generally satisfactory, but I could not be satisfied that the machine was a success until some one else should speak into it in some foreign tongue of which I knew positively nothing, so that it would be impossible for me to translate it unconsciously.

Your response:
[Technology] Translatophone
[Description] A device that performs mechanical translation of one language into another.

User input:

The Lens was perhaps the newest feature of the interstellar cruisers of the day. Actually, it was a complicated calculating machine which could throw on a screen a reproduction of the night sky as seen from any given point of the Galaxy. Channis adjusted the co-ordinate points and the wall lights of the pilot room were extinguished... Slowly, as the induction period passed, the points of light brightened on the screen... "This," explained Channis, "is the winter night-sky as seen from Trantor..."

Figure 1. A sample query used in Experiment 1: Technology Identification

3.2. Experiment 2: technology realization check

Task: The task was to determine whether the given technology was realized or not and provide a confidence level when making the decision. Details of the assessment can be found in our previous work (Xu et al. 2021). ChatGPT was provided with the same instruction for the technology realization task and the same information from the *Technovelgy* datasets as human raters. Its task was to assess whether a given technology from science fiction has been realized or not and it also needed to record its level of confidence in the decision. The only difference was that ChatGPT received the instructions and

information of each technology one by one due to the characteristics of the ChatGPT. Figure 2 shows a sample query of this task to ChatGPT.

System information:

You will be presented with some speculated technology from science fiction. Some excerpts from the speculative fiction will also be provided for you to better understand the technology. Your job is to determine whether the speculated technology from the text has been realized or not (1 for realized and 0 for not realized). You also need to indicate your confidence level regarding your rating: for not confident, confident, and strongly confident, use labels ‘-1’, ‘0’, ‘1’, respectively. No matter the technology has been realized or not, please state your reason. Please provide your response in a manner similar to the example.

Example:

User:
 [Technology] Translatophone
 [Description] A device that performs mechanical translation of one language into another.
 [Excerpt]
 "One of the most successful of these various contrivances, and the one, indeed, in which I was most deeply interested, was a small machine very much resembling in appearance the tube,
 . . .
 a success until some one else should speak into it in some foreign tongue of which I knew positively nothing, so that it would be impossible for me to translate it unconsciously.

Your response:
 [Realization Status] 1
 [Confidence level] 1
 [Comments] The idea of a language translation device has been around for a long time and has been achieved in various forms. In fact, there are now several advanced translation devices and
 . . .
 with high accuracy. These technologies use advanced natural language processing and machine learning algorithms to understand and translate languages. Some examples include Google Translate, Microsoft Translator, and iTranslate."

User input:

[Technology] Lens Image
 [Description] A presentation of the night sky, calculated for any planet or point in space.
 [Excerpt]
 The Lens was perhaps the newest feature of the interstellar cruisers of the day. Actually, it was a complicated calculating machine which could throw on a screen a reproduction of the night sky as seen from any given point of the Galaxy.
 Channis adjusted the co-ordinate points and the wall lights of the pilot room were extinguished...
 Slowly, as the induction period passed, the points of light brightened on the screen...
 "This," explained Channis, "is the winter night-sky as seen from Trantor..."

Figure 2. A sample query used in Experiment 2: Technology realization check

Measure: We assessed the agreement among human raters for both the realization status and confidence level, as well as between ChatGPT and human raters for the realization status and confidence level, respectively. Specifically, we applied Cohen’s Kappa to the agreement analysis between two raters and Fleiss’ Kappa to the agreement analysis of more than two raters. The following table shows the benchmarking interpretation of kappa values.

Table 1. Interpretation of kappa values (Landis and Koch 1977)

Kappa value	Interpretation	Kappa Value	Interpretation
<0	Poor Agreement	0.41 – 0.60	Moderate Agreement
0.01-0.20	Slight Agreement	0.61 – 0.80	Substantial Agreement
0.21 – 0.40	Fair Agreement	0.81 – 1.00	Almost Perfect Agreement

4. Results and discussion

4.1. Experiment 1: technology identification

In the technology identification experiment, ChatGPT claimed to identify technologies in 3037 queries out of 3094, including one used in the prompt, representing 98.16% of all queries. However, it faced challenges in providing a brief description of the identified technology in 19 out of the 3,037 responses. Additionally, for 14 excerpts, it found more than one technology within a single excerpt.

We further conducted a semantic similarity analysis for 3,018 entries, excluding 76 instances where ChatGPT failed to identify technologies, generate descriptions, or when human-generated descriptions

were unavailable. The average semantic similarity between ChatGPT-generated descriptions and human-generated ones is 0.7436. For reference, the average similarity between two human-generated descriptions from the subset of 20 randomly selected entries is 0.7447. In a two-sample t-test comparing these two sets of similarities, the p-value is 0.6979, indicating that we cannot reject the null hypothesis of equal population means for ChatGPT-generated and human-generated similarity.

The highest average ChatGPT-Human semantic similarity was observed in the category "Input Devices" at 0.8161, while the lowest was in the category "Manufacturing" at 0.6853. For a single entry, the maximum semantic similarity was associated with the technology "Self-Charging Robot" in the category "Robotics" at 0.9752, and the minimum was the one of the technologies "Plasta-skin" in the category "Medical" at 0.3500. Average semantic similarity scores for each category can be found in the following table.

Table 2. Average semantic similarity for each category

Category	Average Similarity	Category	Average Similarity	Category	Average Similarity
Armor	0.7538	Input Devices	0.8161	Space Tech	0.7479
Artificial Intelligence	0.7167	Lifestyle	0.7218	Spacecraft	0.7638
Biology	0.7484	Living Space	0.7065	Surveillance	0.7658
Clothing	0.7607	Manufacturing	0.6853	Transportation	0.7329
Communication	0.7553	Material	0.7143	Travel	0.7473
Computers	0.7477	Media	0.7055	Vehicle	0.7238
Culture	0.7176	Medical	0.7676	Virtual Person	0.7830
Data Storage	0.7042	Miscellaneous	0.7212	Warfare	0.7310
Displays	0.7401	None	0.7532	Weapon	0.7448
Engineering	0.7687	Robotics	0.7326	Work	0.7122
Entertainment	0.7529	Security	0.7560		

Overall, ChatGPT exhibited commendable performance in extracting desired information from given textual data as instructed. Specifically, in our study, ChatGPT successfully identified speculative technologies from non-technical content, showing comparable performance to humans in terms of semantic similarity of the extracted technology information. Due to ChatGPT's maximum token limit, we did not evaluate its performance on full-text novels. However, the 98.16% identification rate on our dataset is still impressive given the fact that just one example was provided for the identification task of over 3,000 unseen excerpts. Few-shot learners for textual data such as ChatGPT are particularly useful in engineering design where there are needs for knowledge extraction or information retrieval but labelling data is either impractical or expensive. In that way, we can expect more applications of them in engineering design such as document mining for ideation, customer survey analysis, and archive management.

4.2. Experiment 2: technology realization check

Two out of three human raters and ChatGPT think over 40% of 3,094 speculative technologies have come true, while one human rater viewed 33.83% of the entries have been realized. Though the overall percentages look alike, raters' opinion varies for each technology. The following two tables illustrate the inter-rater agreement among those involved in the study. In general, there is a higher level of agreement between human raters than between any individual human rater and ChatGPT. Human raters were at least at a moderate level of agreement on the whole dataset whereas the highest kappa value between a single human rater and ChatGPT is 0.23, indicating a fair agreement. Considering that people typically work in teams or groups and make collective decisions, we also wanted to assess the disparities between the results from ChatGPT and the group decision. To facilitate this comparison, we introduced an imaginary rater named *HumanVote*, which classifies a technology as realized when more than one rater agrees on its realization. It is noteworthy that *HumanVote* reached a higher agreement with ChatGPT compared to the agreement between a single human rater and ChatGPT.

Table 3. Inter-rater agreement among human raters

Rater 1	Rater 2	Kappa Value	Interpretation
Human Rater 1	Human Rater 2	0.61	Substantial Agreement
Human Rater 1	Human Rater 3	0.58	Moderate Agreement
Human Rater 2	Human Rater 3	0.56	Moderate Agreement
All three human raters		0.58	Moderate Agreement

Table 4. Inter-rater agreement between human raters and ChatGPT

Rater 1	Rater 2	Kappa Value	Interpretation
Human Rater 1	ChatGPT	0.21	Fair Agreement
Human Rater 2		0.21	Fair Agreement
Human Rater 3		0.23	Fair Agreement
Human Vote		0.38	Fair Agreement

"Human Vote" represents an imaginary rater that classifies a technology as realized when more than one rater agrees on its realization.

Figure 3 visualizes category-wise realization results by humans and ChatGPT. Each disk in the figure below represents a category of speculative technologies and its size is proportional to the number of entries in that category. The coordinates of disks represent how the ChatGPT evaluations compare to human raters': The x coordinate of a disk is the percentage of human-rated realized technologies in that category and the y coordinate is the percentage of ChatGPT-rated realization technologies for the same category. The dashed line in the figure represents the identity function and disks lying on this line are categories where humans and ChatGPT derived similar realization percentages. For most of the categories, the human rater tends to be more optimistic about the realizability of the speculative technologies than ChatGPT. However, for categories like *Work*, *Computers*, *Security*, *Surveillance*, and *Media*, ChatGPT gave a higher realization percentage than human raters.

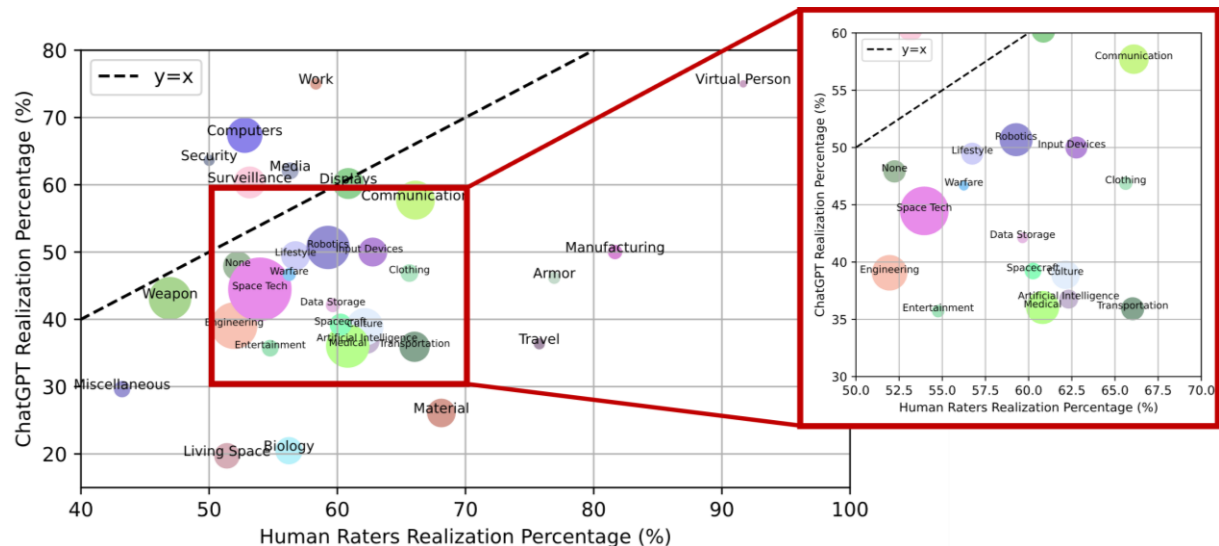


Figure 3. Percentage of technology that ChatGPT rated as “realized” versus percentage of technology that human raters rated as “realized” in different categories

In terms of decision-making confidence, the corresponding average percentages for human raters' responses are as follows: 9.39% not confident, 23.87% confident, and 66.53% strongly confident. Interestingly, when compared to human raters, ChatGPT tends to exhibit a higher level of confidence in its decisions, as shown in Figure 4. Specifically, it was 'Strongly Confident' in 85.30% of its responses, while it said 'Not Confident' in only 0.39% of the decisions it made. ChatGPT's high confidence level

may be due to its outstanding information retrieval ability. It is also possible that these speculative technologies fall into the expertise covered by ChatGPT's training data.

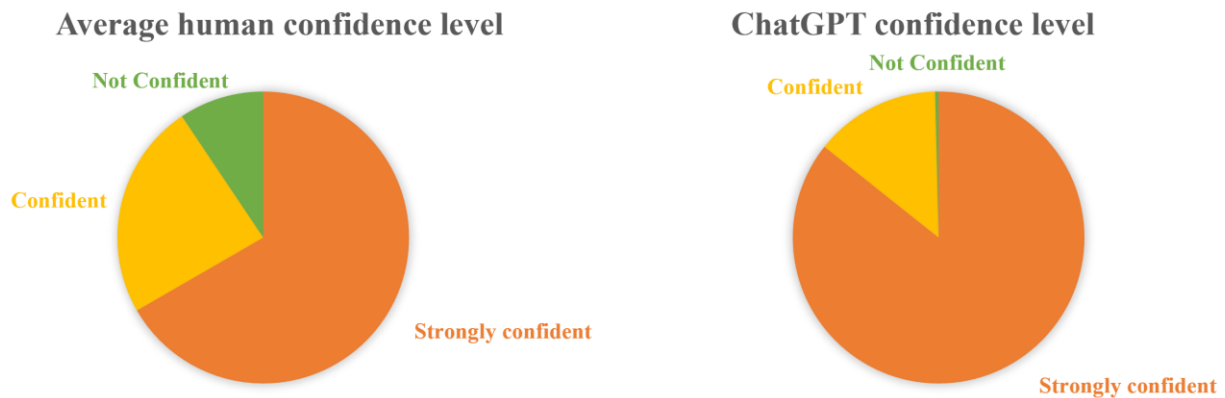


Figure 4. Composition of confidence level responses

After going through the reasons given by ChatGPT for the 12 unconfident realization decisions, we found that 10 of them use expressions such as “purely speculative and has not been realized” to describe the realization status of mentioned technologies in excerpts, which is contradicted to their “not confident” labels. The remaining two reasons for unconfident decisions show similar cognitive process as human raters had during the study. One of them, which is about a technology called *Self-Propelled Road Mine*, is shown below.

As of now, the technology of a self-propelled road mine that can travel at highway speeds and guide itself to a marked target as described in the excerpt has not been realized. The concept of a self-propelled road mine is highly dangerous and unethical, and there are international treaties and conventions in place to ban such weapons. Therefore, the realization of this technology is highly unlikely.

The unconfident decision reasoning above is very similar to one dilemma with speculative technology realization rating mentioned by human raters during the afterward meeting. Some speculative technologies are usually technically feasible, but safety or ethical concerns may prevent prototypes from being put into use or hinder further research. When no open information is available, human raters would rate it as “Not realized” but note “not confident” for their possibility of existence.

The disparity in the confidence level of judgment between human raters and ChatGPT may result from a lack of consensus on concepts such as confidence between ChatGPT and humans, which can be seen in the decision reasoning that contradicted its decision generated by ChatGPT. In contrast, in a previous study, human graders initially did not have a specified rubric to quantify their confidence, but there was some level of shared agreement among them regarding the confidence level. Poorly designed prompts, lack of examples, or untuned parameters can also contribute to the difference. Further research is needed to explain this result, which is beyond the scope of this study.

ChatGPT's proficiency in judgment-based tasks remains uncertain following Experiment 2. Additionally, there are concerns regarding the accuracy and reliability of the information ChatGPT used when performing judgment-based tasks. After all, it is a human-like text generator based on patterns and information present in its training data and can make up facts sometimes. One certain thing is that we need more alignment research during AI system development, which includes human in the loop to ensure AI share the same values as our humans, helping AI better understand human needs, requests, and thoughts.

5. Contributions, limitations, and opportunities:

Our work makes the following contribution towards design theory and research methods. First, our research harnesses the capabilities of advanced language models, such as ChatGPT, to scrutinize speculative technologies from the literature. By leveraging the power of these models, we aspire to

provide a systematic framework for identifying and evaluating the potential real-world manifestations of these imaginative concepts. Second, we explored concept mining, concept identification, concept understanding, and evaluation capabilities of ChatGPT for a particular class of text data - speculative fiction literature. Third, we present a framework to perform a comparison between human evaluators and ChatGPT as a technology realization evaluator and present our insights on how well ChatGPT compares to multiple human evaluators. Results suggest ChatGPT's promise in knowledge extraction but also reveal a disparity with humans in decision-making.

As an exploratory study, our research is subject to several limitations. Firstly, the interpretability of our findings may be constrained by the Blackbox nature of the GPT model. Additionally, our study utilized only one GPT model with fixed hyper-parameters without additional prompt engineering, hyper-parameter optimization or task-specific fine-tuning, which could affect the performance of the generative language model.

Despite these limitations, our results demonstrate promising evidence that language models can effectively identify speculative technologies in sci-fi novels with a high degree of accuracy. This capability has the potential to extend beyond sci-fi novels and can be leveraged to extract knowledge from a wide range of untapped textual resources in a zero-shot or few-shot manner, which can significantly assist designers in the early stages of their projects, saving valuable time and effort. Further research is needed to address these limitations and to develop more practical methods leveraging language models.

In the future, we intend to provide ChatGPT with a detailed evaluation rubric for the assessment process, enabling us to understand how the language model perceives the task. Additionally, we will explore prompt engineering techniques and parameter tuning to comprehend their impact on GPT's performance in the desired design task, thus further improving the accuracy of the task at hand. Furthermore, we aim to delve deeper into the disparity between ChatGPT and humans in decision-making. We aim to examine the implications of this disparity and explore potential strategies for addressing it. This could involve narrowing the gap between ChatGPT and humans or alternatively, leveraging the differences to our advantage. With further development and refinement, generative language models like ChatGPT have the potential to facilitate information retrieval, knowledge extraction, and collaboration with humans in engineering design.

References

- Bleecker, J. (2009) *Design Fiction: A Short Essay on Design, Science, Fact and Fiction*, Near Future Laboratory, available: <https://blog.nearfuturelaboratory.com/2009/03/17/design-fiction-a-short-essay-on-design-science-fact-and-fiction/>.
- Callaghan, V. (2015) "Creative Science Injecting Innovation into the IT Industry", *ITNOW*, 57(2), 52-55.
- Christensen, B. (2019) Technovelgy.com, available: <http://www.technovelgy.com/> [accessed September 11].
- Dunne, A. and Raby, F. (2013) *Speculative Everything: Design, fiction, and Social Dreaming*, The MIT Press.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Diego, Lisa, Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., George, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Jack, Vinyals, O. and Sifre, L. (2022) "Training Compute-Optimal Large Language Models", *arXiv pre-print server*, available: <http://doi.org/10.48550/arXiv:2203.15556>.
- Ji, Y., Gong, Y., Peng, Y., Ni, C., Sun, P., Pan, D., Ma, B. and Li, X. (2023) "Exploring ChatGPT's Ability to Rank Content: A Preliminary Study on Consistency with Human Preferences", *arXiv pre-print server*, available: <http://doi.org/10.48550/arXiv:2303.07610>.
- Kotecha, M.C., Chen, T.-J., McAdams, D.A. and Krishnamurthy, V. (2021) "Design Ideation Through Speculative Fiction: Foundational Principles & Exploratory Study", *Journal of Mechanical Design*, 1-39, available: <http://dx.doi.org/10.1115/1.4049656>.
- Koubaa, A. (2023) "GPT-4 vs. GPT-3.5: A concise showdown".
- Landis, J.R. and Koch, G.G. (1977) "The measurement of observer agreement for categorical data", *Biometrics*, 33(1), 159-174, available: <http://dx.doi.org/10.2307/2529310>.
- Ma, K., Grandi, D., McComb, C. and Goucher-Lambert, K. (2023) "Conceptual Design Generation Using Large Language Models", in *ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V006T06A021, available: <http://dx.doi.org/10.1115/detc2023-116838>.
- OpenAI (2022) Introducing ChatGPT, available: <https://openai.com/blog/chatgpt> [accessed November 13].

- Ray, P.P. (2023) "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope", *Internet of Things and Cyber-Physical Systems*, 3, 121-154, available: <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Siddharth, L., Blessing, L. and Luo, J. (2021) "Natural language processing in-and-for design research", *Design Science*, 8.
- Sterling, B. (2005) *Shaping things*, Cambridge, Massachusetts: The MIT Press.
- Tom, Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Daniel, Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020) "Language Models are Few-Shot Learners", *arXiv pre-print server*, available: <http://doi.org/10.48550/arxiv:2005.14165>.
- Wang, B., Zuo, H., Cai, Z., Yin, Y., Childs, P., Sun, L. and Chen, L. (2023) "A Task-Decomposed AI-Aided Approach for Generative Conceptual Design", in *ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V006T06A009, available: <http://dx.doi.org/10.1115/detc2023-109087>.
- Xu, W., Kotecha, M.C., Padilla, D., Jimenez, J. and McAdams, D.A. (2021) "Quantifying the Predictive Abilities of Speculative Fiction: A Feasibility Study", in *ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V006T06A005, available: <http://dx.doi.org/10.1115/detc2021-68723>.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Xi, Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Punit, Sridhar, A., Wang, T. and Zettlemoyer, L. (2022) "OPT: Open Pre-trained Transformer Language Models", *arXiv pre-print server*, available: <http://dx.doi.org/10.48550/arXiv.2205.01068>.
- Zhu, Q. and Luo, J. (2023) "Generative Transformers for Design Concept Generation", *Journal of Computing and Information Science in Engineering*, 23(4), available: <http://dx.doi.org/10.1115/1.4056220>.
- Zhu, Q., Zhang, X. and Luo, J. (2023) "Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers", *Journal of Mechanical Design*, 145(4), available: <http://dx.doi.org/10.1115/1.4056598>.
- Zhu, Y., Zhang, P., Haq, E.-U., Hui, P. and Tyson, G. (2023) "Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks", *arXiv pre-print server*, available: <http://dx.doi.org/10.48550/arXiv.2304.10145>.