

Strategies in Design Research for Acquiring Scenarios Using Generative AI

ARNGEIR BERGE, NORCE Norwegian Research Centre, Norway

FRODE GURIBYE, University of Bergen, Norway

1 INTRODUCTION

In our DIS'23 paper “Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls” [1], we describe a design research process that could have been sped up had ChatGPT been launched earlier.

The domain of the design research was telephone triage—the kind of medical assessments that are done by telenurses when people call about urgencies to what in the US is known as 911 or Nurse Advice Lines. We prototyped AI support for telenurses located in Norway. The first part of the design process had shown potential for making a prototype that suggested tailored questions that nurses could choose to ask a caller. As a telenurse would document answers for such questions, this would feed into recommender system algorithms that in turn would suggest new questions, all the time following the twists and bends of the conversation between caller and telenurse. To make the algorithms accurate, data on the course of historical conversations were needed, and this turned out to be a problem.

Because of restrictive legislation on call center’s audio recordings and documentation about calls, we were not able to use this kind of scenario data. Instead, we enlisted telenurses over time to write fictional scenarios that bore a resemblance to actual calls they encountered in their work. The result was a modest 257 scenarios with an average of twelve answered questions. We used the scenarios to make a working prototype.

For telenurses to engage with and test the prototype, we needed more scenarios that had not been used to train the algorithms. We invited six telenurses for a role-play workshop, where they took turns playing telenurses and callers. A couple of weeks in advance, we had asked them to document anonymous, but plausible scenarios to act out during the workshop. We had announced three sumptuous chocolate prizes for the three most productive nurses in terms of producing scenarios. The result was an average of just below ten scenarios provided per participant. This was enough but illustrates that acquiring scenarios takes an effort.

We have since done preliminary experimentation with ChatGPT to examine whether it can make scenario acquisition in design research easier, and perhaps leave us some chocolate along the way.

2 SCENARIO DESIGN

Our two problems were acquiring scenarios to train the AI of the working prototype and to use during a role-play workshop. To examine ChatGPT’s capabilities to this end, we made a prompt where we chose a scenario similar to those that the workshop participants provided:

```
I'd like you to generate example problem formulations to a Local Emergency Medical Call Center that receives a wide variety of urgencies that are not life-threatening but should perhaps not wait for an appointment during office hours at the family physician's office. Here is a previous example:
```

```
"I'm calling for my two-year-old girl who has been lying in bed and dozing for a couple of hours now this morning. She wakes up if I try to wake her, but she won't get up. She vomited yesterday, and today she's so hot. I don't have a thermometer, but I've never felt that she has been this hot."
```

What would the three next example calls look like?

Example 1:

Caller: "Hello, I'm calling about my elderly father. He's been complaining of severe stomach pain for the past hour. He hasn't eaten anything unusual, but the pain seems to be getting worse. He also feels nauseous and has a history of acid reflux. I'm not sure if I should take him to the emergency room or wait for his regular doctor's appointment next week."

Example 2:

Caller: "Hi, I need some advice regarding my seven-year-old son. He fell off his bike earlier today and scraped his knee. I cleaned the wound and applied an antiseptic, but it looks quite deep and it's still bleeding. I'm not sure if he needs stitches or if I should just continue with basic first aid. Can you help me determine the best course of action?"

Example 3:

Caller: "I'm calling about my husband. He's been experiencing chest pain for the past half an hour. He's in his late 40s and doesn't have a history of heart problems, but the pain is radiating down his left arm and he's feeling a bit short of breath. I'm worried it might be something serious, but I'm not sure if we should call an ambulance or if it's safe to drive him to the nearest urgent care center."

The scenarios above sound plausible, and making plausible answers is one of ChatGPT's strengths. It is correct that chest pain sometimes radiates down an arm and may be related to a shortage of breath. There is, however, a bias. Indeed, chest pain is quite common, but radiating chest pain would likely not be one of your three arbitrarily extracted scenarios. We prompted ChatGPT to assess the degree of urgency for the three examples, based on a widely used triage system: "Which triage level would you assume for each of the three examples according to The Manchester Triage System Telephone Triage and Advice (MTS TTA)?" ChatGPT provided Very Urgent (Orange) for example 1, Less Urgent (Green) for example 2, and Immediate (Red) for example 3. A telenurse we asked said the assessments also sound reasonable, given the brief information available in the scenarios. ChatGPT is apparently not familiar with the urgency profile of Norwegian local emergency medical communication centers and provided scenarios with too high an urgency. A strategy to make scenarios more plausible or representative at the group level is needed.

In our case, we have data on an overarching level on the characteristics of emergency medical calls. We have a dataset with only the metadata for 1 million calls on age and gender, day and time, urgency level, and healthcare responses. The principal problem of each call in the data set has been coded with one of about 300 symptom codes from the classification system ICPC-2. Such available data can be used to qualify prompts to ChatGPT to make scenarios that together are more representative of calls at large.

When using machine learning on real data to make AI support for calls with real patients, bias attenuation is paramount. For the synthetic scenarios provided by ChatGPT, there will probably be hidden biases. A challenge with this is that prototyping and testing may shape the product to work best towards users and use cases that are most represented. So even when using generative AI to provide scenarios to train prototype algorithms, it is important to use knowledge about the representation of different parts of the patient population to attenuate bias. In reality, highly urgent scenarios are underrepresented compared to how important they are to identify. One would therefore want to

boost their representation so that a prototype is inclined to see scenarios as potentially highly urgent. It is important to strike a balance so that telenurses get help detecting those scenarios without feeling that the prototype cries “Wolf! Wolf!” or “Emergency! Emergency!” for no reason. Even when training algorithms with real data, generative AI may provide a way to boost urgent scenarios by providing plausible synthetic ones. ChatGPT can be prompted from Python code through the OpenAI API so that one can make prompts programmatically and qualify the prompts according to statistical knowledge about real-world scenarios. This can be a way to scale scenario acquisition and quickly design working prototypes for AI support.

In our role-play workshop, the participants expanded on brief scenarios like the ones ChatGPT provided above, using their own experience to provide answers to the telenurse for details not addressed in the scenarios. ChatGPT appears to be able to provide scenarios for this use, though bias attenuation on the group level is needed. It can also generate scenarios that none of the participants are familiar with—something that could be useful in some design research projects.

Regarding scenario acquisition for training machine learning models for design research AI prototypes, the data needs to be structured. In our mentioned study, we had made a terminology of questions and answers that figure in emergency medical calls. To see if ChatGPT could translate a scenario into structured data we made a prompt that also provided relevant structured questions from the terminology:

```
Someone makes the following medical call: "I'm calling for my two-year-old girl who has been lying in bed and dozing for a couple of hours now this morning. She wakes up if I try to wake her, but she won't get up. She vomited yesterday, and today she's so hot. I don't have a thermometer, but I've never felt that she has been this hot."
```

```
Below is a bulleted list with documentation of the call. Can you give me back the exact same list where you simply have chosen one of the answer options (and filled in the blank, _, as needed) based on the call?
```

```
* Caller: [Parent / Partner / Relative / Bystander / Patient]
* Patient age: [_ month(s) / _ year(s)]
* Vomited: [Yes / No / Unknown]
    * Last time: [_ day(s) ago / Unknown]
* Responsive: [Yes / No / Unknown]
* Fever: [Yes / No / Unknown]
    * Has taken the temperature: [Yes / No / Unknown]
* Lethargic: [Yes / No / Unknown]
* Posture: [Lying down / Sitting / Standing / Unknown]
```

```
* Caller: Parent
* Patient age: 2 years
* Vomited: Yes
    * Last time: yesterday
* Responsive: Yes
* Fever: Yes
    * Has taken the temperature: No
* Lethargic: Yes
```

* Posture: Lying down

We see that ChatGPT manages to pick the right answer from the answer options. The formatting is also good, except the answer “yesterday”, which should have been “1 day(s) ago”. Dealing with formatting errors would require automatic and occasionally manual post-processing. We tried adding questions that were not addressed in the scenario, and ChatGPT correctly selected the answer option “Unknown”. This indicates that ChatGPT can do the translation process from unstructured to structured data pretty well, and we see an opportunity to try to scale the effort programmatically so that ChatGPT can provide a large amount of structured data stemming from plausible scenarios.

For scenarios that should be used to train algorithms that can tailor suggested questions during conversations, one should not only know the questions answered but the order in which they were answered as well. The prompt “Can you order each level of items in the bulleted list according to what is uncovered first and last in the call?” gave a pretty good order of the list; we agreed on the placement on seven out of nine points.

3 CONCLUSION

In a design research effort for our DIS'23 paper, our main problems were to acquire scenarios for training machine learning models for prototyping AI interaction and for use in a role-play workshop. After the launch of ChatGPT, we have done a preliminary examination of whether ChatGPT could alleviate these problems for future design research projects.

The scenarios that ChatGPT provided for prototype testing were plausible. Also, ChatGPT seems able to transform scenarios into synthetic data that can be used to speed up the design and training of working prototypes of AI interaction. The provided scenarios were biased towards high urgency, but since ChatGPT was able to map out the urgency of the scenarios it provided, a strategy may be to use ChatGPT programmatically to qualify prompts with real-world metadata so that bias can be attenuated, and to lower the risk of attuning the design away from serving underrepresented groups and health conditions.

A challenge with prototyping interaction with AI is the time it takes to acquire data to make a working prototype. Data acquisition in the health area can, if possible in the given country, take months, and may require several rounds of clarifications back and forth between approval committees and researchers. Also, it can be time-consuming to make scenarios for testing a working prototype. Designing believable scenarios can require expertise in a domain that the researchers are not familiar with. Our examinations indicate that ChatGPT can speed up design research efforts by providing scenarios as training data and for role-play workshops to study human-AI interaction.

REFERENCES

- [1] Arngeir Berge, Frode Guribye, Siri-Linn Schmidt Fotland, Gro Fønnes, Ingrid Hjulstad Johansen, and Christoph Trattner. 2023. Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls. In *Proceedings of ACM DIS'23*. ACM, Pittsburgh, PA, 14.