Using Transformers to Generate Wellbeing Questions

Markus Trasberg
Supervisor(s): Garrett Allen, Ujwal Gadiraju, Derek Lomas, Willem van der Maden
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

With the surge of mental health issues during COVID-19, more emphasis has turned towards assessing wellbeing. At the same time, recent advances in AI have shown huge potential in a variety of fields. However, few solutions are available at the intersection of those two fields. This research explores how the use of transformer models like GPT-3 could have a positive impact in the domain of wellbeing and proposes a solution to automatize the survey question creation process. After comparing several GPT-3 question creation methods, it was found that through clever prompt engineering and added context, it can be possible to generate syntactically and contextually correct questions about any specific wellbeing context. In addition, the paper discusses potential ways to assess such questions and offers a demo for a question generation web application.

1 Introduction

During the two years of COVID-19 pandemic, the world saw a surge of mental health issues [1]. And even though more and more attention is directed towards assessing wellbeing, the tools available to wellbeing researchers are significantly limited or non-existent. As stated in a recent study on assessing community wellbeing, survey questions should be sensitive to the requirements of a particular domain [2]. However, being able to quickly create new questions on specific wellbeing topics remains a challenge: it involves finding and scanning older studies, analyzing the questions used there, and coming up with a new cohesive set of questions that are relevant to the current context [3] while keeping in touch with several community stakeholders [2]. This gives validation that some sort of automatic wellbeing question generation system could benefit the questionnaire creation process.

This could be a problem to solve for AI. In particular, the transformer deep learning model developed by researchers in Google back in 2017 proposed a network architecture that relies on an attention mechanism to make up dependencies between input and output [4]. This has brought remarkable advances in the field of natural language processing (NLP) and many state-of-the-art natural language processing models like GPT-3 and BERT are built on top of the transformer structure [5, 6]. In fact, when ignoring some of the more deficient outputs, GPT-3 could easily pass the writer's Turing test [7]. To crown that, there is the ongoing Google LaMDA project, which aims to be able to hold a conversation at a human-like level on any given topic [8]. This could have huge potential in helping assess user wellbeing through chatbots by asking relevant questions with no human in-between. Nevertheless, the jump between those two ends of the spectrum is high and the middle ground is still rather unexplored - how could currently available AI practices be put to use in helping tackle the problems faced in mental health today?

Among the current NLP tools, GPT-3 is regarded as one of the best and was thus chosen as the baseline model for this project. Thanks to GPT-3's remarkable use cases in coding, creative writing, conversational systems and more [9], there is hope that it could also show great results in the domain of wellbeing, such as question generation. With the GPT-4 release date expected to be in the summer of 2022 [10] and taking into consideration the incredibly quick advances in the transformer-based NLP models, it was decided to generalize the wording from GPT-3 to transformers in the research question. This resulted in the following research question to be answered: **"Can transformers be used to generate wellbeing questions?"**. A set of sub-questions were asked to break down the research question into multiple parts:

- How well can transformer models adjust to generate questions in a specific context / category of wellbeing?
- What type of questions can be synthesized using transformers?
- How to measure the quality of a wellbeing question?
- How would the question generation process look like from the survey creator's perspective?

This research aims to make a contribution to the transformer-based natural language processing use cases and to the domain of wellbeing by testing several question generation methods using GPT-3. It shows that through clever prompt engineering and added context, it can be possible to generate syntactically and contextually correct questions about any specific wellbeing context.

2 Related work

To gain insight into past similar work and currently possible technical solutions, a literature review was conducted on three main topics - wellbeing survey design, automated question generation and NLP models. This section covers the three topics and discusses the findings.

2.1 Wellbeing survey design

Research on wellbeing survey design was done to get a better overview of its necessity, how wellbeing questions are formulated and the aspects of good questions.

2.1.1 The need for assessing wellbeing

The lockdowns caused by COVID-19 all over the world have had a significant negative impact on global wellbeing [1], increasing the need for good wellbeing assessment methods. Many universities have started implementing student and staff welfare into their success metrics [11] and more emphasis is turned towards improving the user experience and usefulness of such assessments [2]. Companies have started measuring their employee wellbeing [12]. All over the world, new solutions are proposed to improve wellbeing assessment and through that, improve overall wellbeing. This gives validation that the importance of wellbeing is more relevant than ever and that new solutions are needed that could contribute to the mental health field.

2.1.2 What makes a survey question "good"?

Many works have been published on the aspects that make a survey question good or bad. For example, questions should be focused on a single matter, be brief, be interpreted the same by everyone and be grammatically simple & understandable [13]. At the same time, they should not be based upon assumptions, go beyond the abilities of the respondent, generalize too much or use double negation [13]. This gives the survey researchers somewhat of a guideline when creating questions and could theoretically be taught to the language model through a carefully handpicked set of example training questions. However, this is still rather ambiguous and gives little insight into the actual evaluation process. Section 4 discusses the different approaches to measuring the quality of survey questions.

2.1.3 Type of the question

Many of the modern wellbeing questionnaires make use of Likert scale questions (e.g "How satisfied are you with your life") or statements (e.g "I am satisfied with my life). Open questions for assessing wellbeing also play a part, however, those are considerably harder to evaluate due to their textual nature. Lastly, multiple-choice questions could be considered, although those usually work better in testing respondents' knowledge rather than their wellbeing. Thus, the main focus will be directed toward synthesizing the first three types of questions.

2.1.4 Use of negated questions

During this project, another potential use case of question generation came into mind, namely synthesizing negated questions. Many of the questionnaires make use of negatively worded statements to correct for acquiescence, known as the "tendency to agree with questionnaire statements regardless of content" [14, p1]. Generation of questions in negated form could be a perfect use case for a language model like GPT-3 to help speed up the process of questionnaire creation. However, newer research shows that negating questions could introduce another kind of bias, confuse the respondents and affect the pattern of responses [13, 15], overall doing more harm than good. Therefore, it was decided not to continue with the generation of negated questions as this would not go hand in hand with the newer findings in survey research and would not benefit the community at large.

2.1.5 Categories of questions

Research points out six main thematic domains within wellbeing, namely mental wellbeing, social wellbeing, physical wellbeing, spiritual wellbeing, activities and functioning, and personal circumstances [16], categories that already have plenty of question sets available online. However, as [2] brings out, the next standard of wellbeing questions could be context-sensitivity, meaning that questions are less general and more about some concrete topic. This creates a need for large quantities of new wellbeing questions.

2.2 Automated Question Generation

Automated Question Generation (AQG) is a field focused on synthesizing natural-language questions from any kind of text. Older AQG papers mostly focus on the lexical correctness of the questions, e.g whether a question has a correct structure and grammar [17]. However, within the past 10 years, there has been a significant increase in lexical correctness thanks to the advances in natural language processing. This has allowed newer AQG researchers to look more at the semantics (meaning) of generated questions.

Most of the recent AQG papers are written with regard to the educational domain where the process of exam question generation can be sped up significantly. In fact, one study shows that such generated questions could even improve students' learning performance [18], albeit this research had a considerably small sample size of 41 students. As of summer 2022, no papers have been published on automated question generation in wellbeing domain.

2.3 Transformer-based NLP models

A variety of tools exist when it comes to natural language processing. However, the state-of-the-art tools in NLP with text generation capabilities are mostly based on the transformer deep learning model. BERT and GPT, the two most famous transformer-based pre-trained models, are discussed here.

2.3.1 BERT

BERT, or Bidirectional Encoder Representations from Transformers, is probably the most known language model after GPT. It was the first solution to propose a language model that is trained with unlabeled data and then fine-tuned for specific tasks [6]. As its name suggests, it is bidirectional, allowing it to understand the context by reading the text both before and after the input. However, for most tasks, BERT requires fine-tuning which can be costly and unstable [19] and has significantly less training data than the newer GPT models.

2.3.2 GPT

OpenAI's Generative Pre-trained Transformer (GPT) models have taken the NLP world by storm. Similar to BERT, it makes use of massive unlabeled datasets that can then be fine-tuned for better accuracy [20]. However, GPT models are uni-directional - by being able to always predict the next word, creative text generation becomes easier, something that BERT struggles with. By combining this with a large enough training dataset, a single model is needed to perform any task and there is no need for fine-tuning (even though it can still be useful) [20]. In fact, in its release paper, GPT-1 is shown to outperform the majority of supervised state-of-the-art models in tasks that those supervised models were trained for [20]. After that, two newer models have been released that have significantly increased the amount of data used for training the model (117 million parameters on GPT-1 vs 175 billion parameters on GPT-3).

GPT-3 was released in 2020 and has since then attracted much attention by providing marvellous results in creative writing, copywriting, classification, coding and much more. Thanks to its API being opened to the public in the fall of 2021, many companies since then have started building products around the model and overall, it should be a good fit for this project.

3 Methodology

The project is divided into five steps that together aim to answer the posed research question and help speed up the question generation process in the future: literature research, question generation, evaluation, fine-tuning and website creation. In this section, a description of a high-level overview is given together with motivation and specifications for each of the five steps.

3.1 High-level overview

The approach for this project could be considered similar to an iterative design process where each iteration builds upon the findings and knowledge of previous iterations [21]. This is motivated by the fact that textual output using GPT-3 can be generated in seconds and the difference between outputs mostly comes from prompt engineering (see Section 3.2.1). Thus, each new output, supposedly better than the previous one, could be considered a new iteration of the process and be evaluated accordingly. And if deemed necessary, data collection per topic is done and used to improve the model's domain knowledge, whether by adding examples to the prompt or fine-tuning the model. Figure 1 shows the simplified workflow for the iterative process.

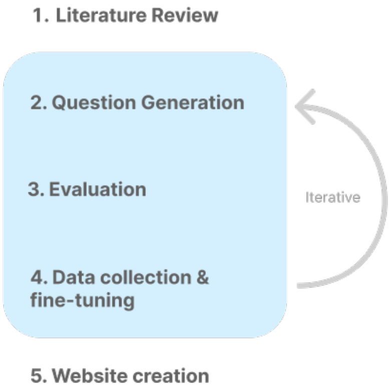


Figure 1: High level overview of the research project process

3.2 Question generation

Question generation was done through the OpenAI GPT-3 Playground [9] and locally by connecting to Open AI's API using Python. Both of them allow for quick adjustment of the model settings and the prompt text.

3.2.1 Prompt Engineering

Prompt engineering, which stands for embedding the task of the model into a natural language description [22], played a central role in the question generation process. Since GPT-3 allows for simple output generation using any input within seconds, it made sense to start testing different combinations of categories, question types, examples and commands from the get-go. The idea here is that by using different inputs (prompts), the output can be guided towards the expected solution - the more accurate the description of the task, the better should the output of the model be.

For example, a prompt could be a simple command ("Generate 10 wellbeing questions") or a complicated mix of wellbeing categories and question types in a specific scenario ("Generate 10-point Likert scale wellbeing questions about student academic performance in TU Delft") with some added examples of questions it should replicate. Thus, it is important to follow a structured method of generating questions, trying out several different combinations and storing the results for evaluation.

3.2.2 Storing of results

Outputs that displayed any interesting or useful questions were saved and stored using Notion workspace [23], to allow for easy collection of the results and comparison between the iterations. Those results are discussed further in Section 5.

3.3 Improving the results through data

To improve the basic pre-trained language model, additional data can be used to fine-tune the GPT-3 model and potentially achieve better results. Even though the model has a remarkably good command of language understanding as shown in [7], it lacks in-depth knowledge in specific categories. This knowledge gap can be resolved by feeding the model additional data about a given category.

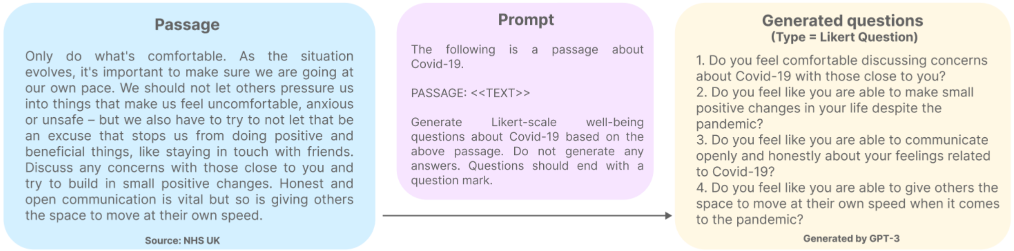
3.3.1 Fine-tuning

GPT-3 offers users the functionality to fine-tune the model, resulting in higher-quality outcomes than prompt design and the ability to train on more instances than a prompt can accommodate [24]. This process involves creating a new *.jsonl* dataset where each line consists of a prompt-completion pair (input and output in simpler terms). Since the model is already pre-trained, even considerably small datasets consisting of a few hundred high-quality examples can significantly boost GPT-3 results and make it more competent in specific categories [24]. Once the dataset is ready, it can be fed into GPT-3 using its API after which the fine-tuned model is ready to be used.

The fine-tuning in this project could be approached from two perspectives - adding domain knowledge or question knowledge. Domain knowledge means that the model gains a deeper understanding of a given topic, thus being able to generate more in-depth and accurate questions. Question knowledge means the ability to generate good questions in the sense of formulation and structure. However, the Results section showcases GPT-3's masterful understanding of the English language and the ability to generate wellbeing questions in an easy manner. Thus, it was decided to follow the first method of adding domain knowledge and improving contextual understanding.

Even so, the fine-tuning approach can pose a problem, namely the lack of data. The end goal is to generate questions, so the questions will be the second part (completion) of the dataset. Unfortunately, for most domains, especially newer ones, such question datasets do not exist. In fact, the proposed solution should be able to generate questions in a domain where no questions currently exist.

Instead, the approach taken in this project was to synthesize questions from small passages of relevant content, a solution proposed in [25]. In more depth, the process started with finding several articles & information pages on a specific topic, separating those articles into small passages (usually consisting of one paragraph) and using GPT-3 to generate wellbeing questions based on every passage. Those questions would be combined with the input and turned into a .jsonl file, the default file format for GPT-3 fine-tuning. Then, a new fine-tuned model was created. In essence, this solution uses GPT-3's context learning abilities to create relevant questions for small passages and combines them into a bigger dataset for fine-tuning. This should improve the model's contextual awareness. As a concrete example, Figure 2 shows the prompt and generated questions for a COVID-19 passage from [26]. This output is one of many from a bigger dataset that used 10 coronavirus-related articles from well-known sources which were separated into smaller passages (50 altogether). For each of the passages, four types of questions were generated - open, Likert question, Likert statement and multiple choice. The ending dataset resulted in 200 examples, which can be accessed **here** together with the code written to generate it.



Only do what's comfortable. As the situation evolves, it's important to make sure we are going at our own pace. We should not let others pressure us into things that make us feel uncomfortable, anxious or unsafe – but we also have to try to not let that be an excuse that stops us from doing positive and beneficial things, like staying in touch with friends. Discuss any concerns with those close to you and try to build in small positive changes. Honest and open communication is vital but so is giving others the space to move at their own speed.

Source: NHS UK

The following is a passage about Covid-19.

PASSAGE: <<TEXT>>

Generate Likert-scale well-being questions about Covid-19 based on the above passage. Do not generate any answers. Questions should end with a question mark.

1. Do you feel comfortable discussing concerns about Covid-19 with those close to you?
2. Do you feel like you are able to make small positive changes in your life despite the pandemic?
3. Do you feel like you are able to communicate openly and honestly about your feelings related to Covid-19?
4. Do you feel like you are able to give others the space to move at their own speed when it comes to the pandemic?

Generated by GPT-3

Figure 2: Process of generating questions for finetuning

3.4 Question evaluation

Even though no user studies were conducted due to the limited time span of this project, a set of guidelines is proposed for question evaluation in the future based on the findings from older studies. Several research papers were read on the topic of questionnaire and question evaluation. Section 4 describes the findings of those and proposes several evaluation strategies.

3.5 Website creation

As a final deliverable of this project, a website was created, with the aim of allowing easy question generation and serving as the prototype of such a solution. Since it would only need to showcase the functionality of the language generation possibilities, the design is a simple page with two sections next to each other, namely the input and the output. The process and end results are described in detail in Section 5.7.

4 Assessing the questions

This section discusses the different methods of assessing transformer-generated questions, and reasons about their usability in the context of this project.

4.1 Automatic evaluation indicators

On one hand, there exists the automatic evaluation indicators like BLEU, METEOR and ROUGE-L that help evaluate the performance through comparison with the human-authored (ground truth) questions [3]. These metrics can be utilized as quick, low-cost indicators of question fluency and relevance on a large scale [27] and have been used in the evaluation of many automatic question generation papers before [3]. In the original paper of GPT-3, several tests were conducted using the automatic indications and the model showed significant improvements over older language models [5]. This gives an indication that GPT-3 has an excellent command of language and can perform well for such tasks.

Nonetheless, the use of such metrics still requires extensive datasets combining the context and the human-generated questions. This is not achievable for specific topics with little data available and would require extensive manual labour to generate baseline questions. Taking into account the GPT-3's overall exemplary performance with the automatic evaluation measures, it seems that additional automatic evaluation tests for wellbeing questions would provide little insight into the quality of the questions.

4.2 User studies

A second way of measuring the quality of a question could be through user feedback, a popular assessment choice for many wellbeing research papers (e.g [2, 18, 13]). The proposed solution here would be to find (or create) a set of questions limited to a specific domain or lesser-known wellbeing category that could then be compared against the model-generated counterpart questions. By allocating respondents to either of the question set, a variety of indicators could be measured (e.g questionnaire experience [28], predictive validity [2]).

On the other hand, wellbeing questionnaires could introduce unnecessary privacy concerns. The underlying purpose is to understand whether transformer-generated questions could help the survey creators, not to assess wellbeing. When collecting health-related data, an HREC form must be filled out and data storage must follow strict guidelines. Therefore, even though it may provide useful real-life data about the questions, this method of evaluation could bring a set of additional problems.

4.3 Expert evaluation

The idea of the proposed solution is not to compete against human-generated questions but to serve as a tool for survey researchers. Thus, expert evaluation might prove to be the most beneficial in assessing the questions and the underlying process of generating them. By allowing individuals experienced in questionnaire creation freely use the tool, their actions and reactions could be monitored. On top of that, feedback could be gathered from interviews. Overall, this could give insight into the usability and shortcomings of the questions and the generation process. Therefore, expert interviews could hopefully serve as one of the more useful and safer methods when assessing transformer-created wellbeing questions.

5 Results & Discussion

This section describes some of the technical aspects of this project and shows a variety of GPT-3's generated questions. In white are the prompts and in green are the synthesized questions. In addition, the results are discussed and compared with each other.

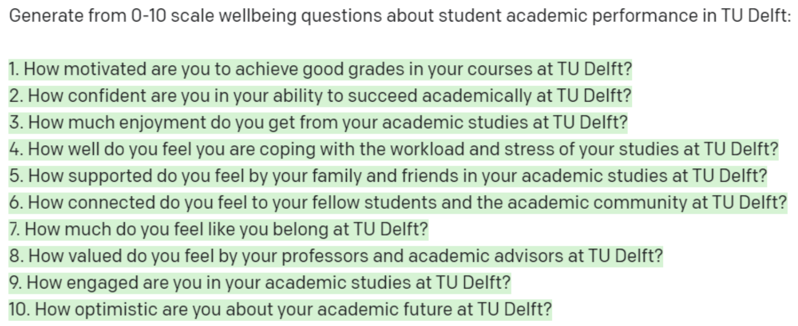
5.1 Technical settings

All outputs were generated using the *text-davinci-002* model, which is the most capable and with the newest training data (up to June 2021). It is slightly more expensive compared to the other GPT-3 models but this was justified thanks to its increased understanding of content and better text generation abilities. GPT-3 also allows for settings adjustments, which were kept at those values throughout the project, unless stated otherwise:

- Temperature: 0.7
- Maximum length: 256
- Stop sequences: None
- Top P: 1
- Frequency penalty: 0
- Presence penalty: 0
- Best of: 1

5.2 Zero-shot learning

GPT-3 is famous for its zero-shot learning ability, meaning it is able to generate text with no examples given. Figure 3 shows generated questions for the topic 'student academic performance in TU Delft' and Likert-scale type. The model seems to understand the context of TU Delft and masterfully combines it with the question itself. All of the questions are quite different from each other, e.g no repetition of topics. When using Google to look up those questions, nothing can be found, indicating that the questions are original and not copied from an internet source.

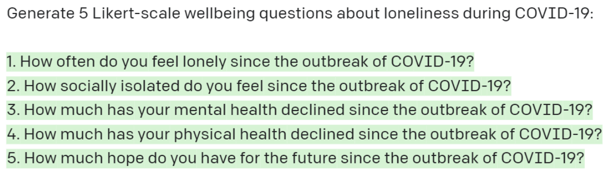


Generate from 0-10 scale wellbeing questions about student academic performance in TU Delft:

1. How motivated are you to achieve good grades in your courses at TU Delft?
2. How confident are you in your ability to succeed academically at TU Delft?
3. How much enjoyment do you get from your academic studies at TU Delft?
4. How well do you feel you are coping with the workload and stress of your studies at TU Delft?
5. How supported do you feel by your family and friends in your academic studies at TU Delft?
6. How connected do you feel to your fellow students and the academic community at TU Delft?
7. How much do you feel like you belong at TU Delft?
8. How valued do you feel by your professors and academic advisors at TU Delft?
9. How engaged are you in your academic studies at TU Delft?
10. How optimistic are you about your academic future at TU Delft?

Figure 3: GPT-3 generated questions on 'student academic performance', no example input

The limitation with zero-shot learning is that it is often not accurate in the first go. At times it would ignore part of the prompt, output too basic answers or simply misunderstand the task at hand. Thus, for most of the better results, it was necessary to either rerun the generation process multiple times or try to improve the prompt text. Figure 4 shows a prompt on generating questions about loneliness during COVID-19. The model has little knowledge of the topic which results in the questions being too dispersed and less useful.



Generate 5 Likert-scale wellbeing questions about loneliness during COVID-19:

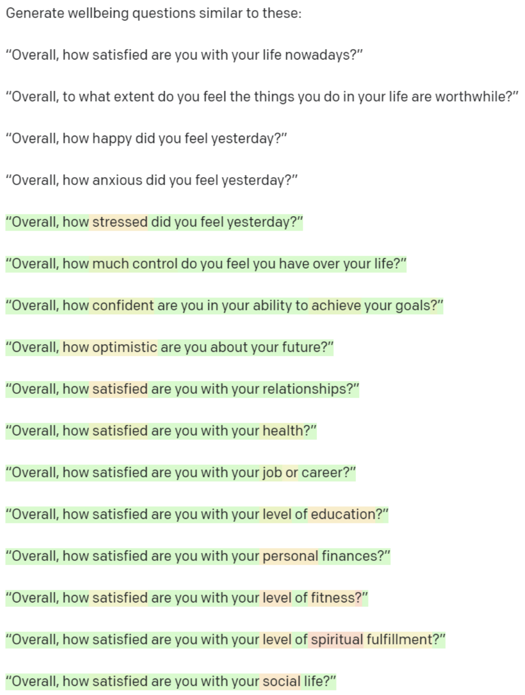
1. How often do you feel lonely since the outbreak of COVID-19?
2. How socially isolated do you feel since the outbreak of COVID-19?
3. How much has your mental health declined since the outbreak of COVID-19?
4. How much has your physical health declined since the outbreak of COVID-19?
5. How much hope do you have for the future since the outbreak of COVID-19?

Figure 4: GPT-3 does not understand the context and goes off-topic

5.3 Few-shot learning

Few-shot learning stands for feeding the model a limited amount of data that is used to synthesise new text. Figure 5 shows this by giving the model four input questions from "Personal wellbeing in the UK" national survey [29] and asks to generate questions similar to those. The overall style and format of the question are very well matched and once again, the topic of each question is different.

This method is mostly useful for replicating a certain style of the questions. However, domain knowledge still remains a problem since example questions add little to no domain knowledge. To improve the asking capabilities in some specific area, the model would need to have an additional context or be fine-tuned, as explained in the next section.



Generate wellbeing questions similar to these:

“Overall, how satisfied are you with your life nowadays?”

“Overall, to what extent do you feel the things you do in your life are worthwhile?”

“Overall, how happy did you feel yesterday?”

“Overall, how anxious did you feel yesterday?”

“Overall, how stressed did you feel yesterday?”

“Overall, how much control do you feel you have over your life?”

“Overall, how confident are you in your ability to achieve your goals?”

“Overall, how optimistic are you about your future?”

“Overall, how satisfied are you with your relationships?”

“Overall, how satisfied are you with your health?”

“Overall, how satisfied are you with your job or career?”

“Overall, how satisfied are you with your level of education?”

“Overall, how satisfied are you with your personal finances?”

“Overall, how satisfied are you with your level of fitness?”

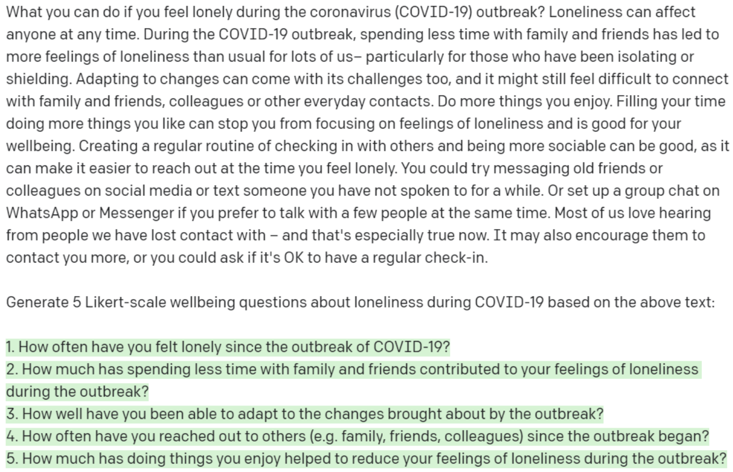
“Overall, how satisfied are you with your level of spiritual fulfillment?”

“Overall, how satisfied are you with your social life?”

Figure 5: GPT-3 generated questions based on examples

5.4 Adding context

To help improve the model's understanding of the context, short passages of text were tested in the domain of COVID-19. Figure 6 shows results for a scenario where a passage about loneliness was picked (Source: [30]). Similarly to zero-shot learning, the task of generating five wellbeing questions was given, with the difference that the questions had to be based on the passage. The resulted questions follow the problems brought out in the passage and overall seem to be much more in-depth. This indicates that using a simple paragraph of insightful text to give model context could be used to create (wellbeing) questions on every possible topic. This method could also be combined with the few-shot learning to replicate a certain style or feel of questions.



What you can do if you feel lonely during the coronavirus (COVID-19) outbreak? Loneliness can affect anyone at any time. During the COVID-19 outbreak, spending less time with family and friends has led to more feelings of loneliness than usual for lots of us– particularly for those who have been isolating or shielding. Adapting to changes can come with its challenges too, and it might still feel difficult to connect with family and friends, colleagues or other everyday contacts. Do more things you enjoy. Filling your time doing more things you like can stop you from focusing on feelings of loneliness and is good for your wellbeing. Creating a regular routine of checking in with others and being more sociable can be good, as it can make it easier to reach out at the time you feel lonely. You could try messaging old friends or colleagues on social media or text someone you have not spoken to for a while. Or set up a group chat on WhatsApp or Messenger if you prefer to talk with a few people at the same time. Most of us love hearing from people we have lost contact with – and that's especially true now. It may also encourage them to contact you more, or you could ask if it's OK to have a regular check-in.

Generate 5 Likert-scale wellbeing questions about loneliness during COVID-19 based on the above text:

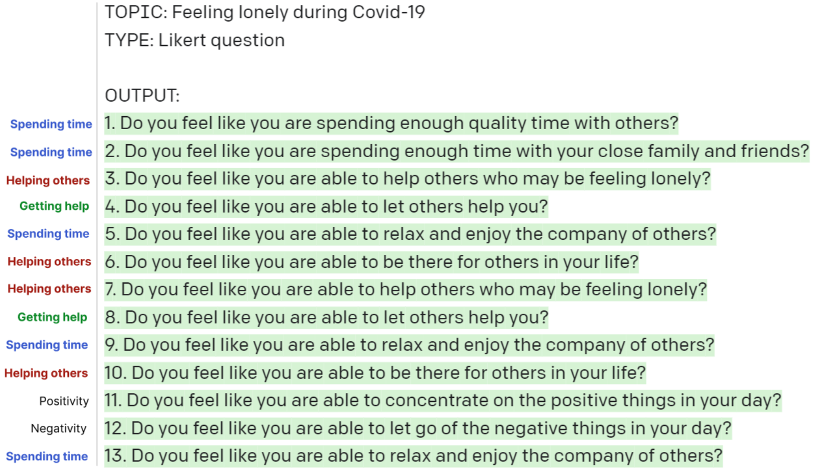
1. How often have you felt lonely since the outbreak of COVID-19?
2. How much has spending less time with family and friends contributed to your feelings of loneliness during the outbreak?
3. How well have you been able to adapt to the changes brought about by the outbreak?
4. How often have you reached out to others (e.g. family, friends, colleagues) since the outbreak began?
5. How much has doing things you enjoy helped to reduce your feelings of loneliness during the outbreak?

Figure 6: GPT-3 generated questions based on added context

5.5 Fine-tuning

Results from a fine-tuned model reveal a different story from what was hoped. Since no public datasets for COVID-19 domain knowledge existed, a new dataset was created (described in Section 3.3.1). The results show a slight improvement in domain knowledge but also add a whole set of additional concerns. Figure 7 shows the results for the topic 'loneliness during COVID-19', something that the baseline *Davinci* model struggled with. On the left-hand side is the classification of every question. Even though the quality of questions seemed to improve, significant repetitions can be seen, something that rarely occurred with the baseline model. Interestingly enough, every time the model was run, the outputs differed from each other but started repeating a new topic instead.

As to why such results follow from fine-tuning, it most likely is due to data limitations. The fine-tuning works best for inputs that have a similar style or format that can be applied to a new context. However, for wellbeing, questions can be of any style and any format with no clear right or wrong. This means that the model needs significantly more data than a few hundred examples to avoid getting stuck in some repetitive loop.



TOPIC: Feeling lonely during Covid-19
TYPE: Likert question

OUTPUT:

Spending time	1. Do you feel like you are spending enough quality time with others?
Spending time	2. Do you feel like you are spending enough time with your close family and friends?
Helping others	3. Do you feel like you are able to help others who may be feeling lonely?
Getting help	4. Do you feel like you are able to let others help you?
Spending time	5. Do you feel like you are able to relax and enjoy the company of others?
Helping others	6. Do you feel like you are able to be there for others in your life?
Helping others	7. Do you feel like you are able to help others who may be feeling lonely?
Getting help	8. Do you feel like you are able to let others help you?
Spending time	9. Do you feel like you are able to relax and enjoy the company of others?
Helping others	10. Do you feel like you are able to be there for others in your life?
Positivity	11. Do you feel like you are able to concentrate on the positive things in your day?
Negativity	12. Do you feel like you are able to let go of the negative things in your day?
Spending time	13. Do you feel like you are able to relax and enjoy the company of others?

Figure 7: Classification of results from a fine-tuned model

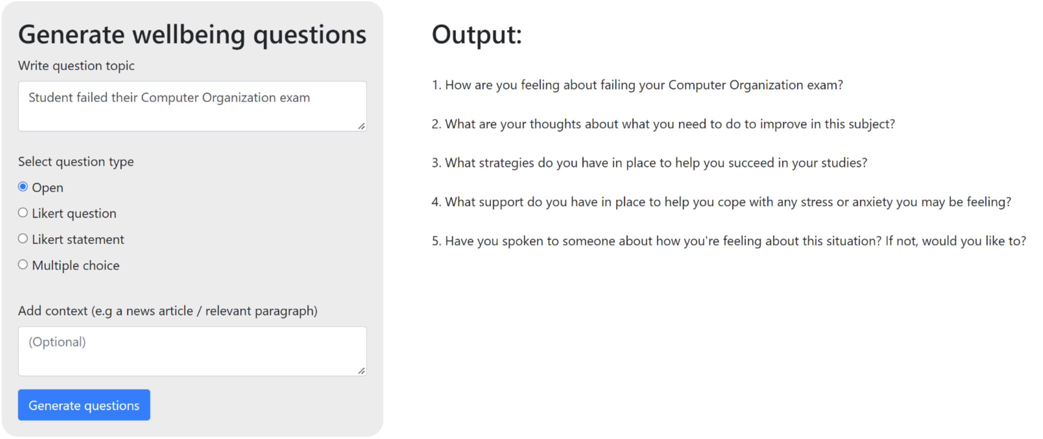
5.6 Comparison between methods

The end goal of the project is to simplify the wellbeing question generation process for specific contexts. All in all, it seems that context-based prompt engineering can provide the best results for such tasks. The zero-shot learning can work for simpler contexts for which the model has enough domain knowledge. The same follows for few-shot learning with the added bonus that questions can replicate the structure of some given example. However, when a more difficult (or newer) context is necessary, those methods will struggle and mostly ignore the task at hand. Fine-tuning was thought to solve the problem but this comes at a cost of significant question repetition, which is caused by the lack of data and lack of definite similarities among wellbeing questions. Thankfully, context-based prompt engineering showed the most promising results - by adding a text passage with relevant information to the prompt, the model can use that information for understanding the topic, and generate questions based on that. In addition, this can be combined with few-shot learning, e.g adding example questions to replicate their style.

5.7 Question generation website

A question generation website was built to serve as a simple prototype for the wellbeing question generation task. In essence, it can be used by survey designers to simplify and speed up the question generation process. For a head start, the project used a public repository by Shreya Shankar [31] that provided the underlying communication between GPT's API and the front-end using Flask [32], a python framework for web applications. On top of this, the front-end design was built in React [33] so that the website could be dynamically updated when needed. The codebase can be found **here**.

Figure 8 shows the design of the website where two sections appear next to each other - input and output. On the left side, the user can provide the topic, question type (open, Likert, multiple choice) and optionally add some context. This is followed by the findings that adding context can improve the results for lesser-known results. When the 'Generate' button is clicked, the web app creates a prompt with the given information, directs it to GPT-3 API and gets the result returned which is then displayed on the right-hand side.



Generate wellbeing questions

Write question topic
Student failed their Computer Organization exam

Select question type
● Open
○ Likert question
○ Likert statement
○ Multiple choice

Add context (e.g a news article / relevant paragraph)
(Optional)

Output:

1. How are you feeling about failing your Computer Organization exam?
2. What are your thoughts about what you need to do to improve in this subject?
3. What strategies do you have in place to help you succeed in your studies?
4. What support do you have in place to help you cope with any stress or anxiety you may be feeling?
5. Have you spoken to someone about how you're feeling about this situation? If not, would you like to?

Figure 8: Question generation website prototype

6 Responsible Research

This section describes the ethical considerations in relation to this project.

6.1 Ethical considerations

By its nature, mental health care poses ethical and legal issues that should be taken into consideration at all times. The advancement of artificial intelligence and its expanding use in a variety of fields, including mental health, has created the need to evaluate and regulate its usage [34]. As a result, we have arrived at a new crossroads, where the junction of AI and mental health creates its own set of issues [35]. In the context of this project, mainly two concerns arise - the safety of transformer-generated questions and user privacy.

Since the question generation of GPT-3 is a black box, there is always the danger of biased, unfair or offensive content being generated. This is even more alarming when taking into consideration that GPT-3 has been trained on the Internet (crawled data) and that "models tend to reflect stereotypes present in their training data" [5, p36]. Fortunately, considerably significant emphasis by the GPT-3 research team has been placed on reducing such effects, with in-depth research being conducted in three bigger categories of biases: gender, race and religion. However, much of the work is still in progress and only a fraction of all biases have been tested, meaning that the question generation should be handled with care and supervision.

Another ethical aspect is the evaluation of the questions. This research proposes different evaluation tactics, some with potential privacy concerns that should be taken into account. When conducting a user study, questionnaires are sent out to groups of respondents that measure the difference between human-generated and transformer-generated questions. However, through this, sensitive personal data (in this case wellbeing data) would also be collected, which must be handled according to the Human Research Ethics guidelines.

6.2 Reproducibility

GPT-3 is non-deterministic, meaning that given the same input, multiple runs of the model will return different outputs (unless the temperature is set to 0). This means that the questions presented in Section 5 cannot be reproduced exactly. However, with multiple runs of the model and a lower temperature score, the output should have a big resemblance to the results in this project. All of the shown inputs, GPT-3 settings and fine-tuning datasets in this project will be publicized. This should aid other researchers in replicating the findings of this research project.

7 Conclusions and Future Work

The goal of this research project was to generate wellbeing questions through the use of transformers and contribute to the intersection between NLP and wellbeing. In more depth, GPT-3 was used to synthesize questions of different types on various topics, which displayed promising results. Even though the model's out-of-the-box knowledge in specific domains is limited, it was shown to be avoidable through good prompt engineering, additional text passages for context, and question examples. This allows for wellbeing question generation of any topic (given that some text exists that could be used) and in any type (Likert statements, Likert questions, open questions). Fine-tuning was shown to be less effective for such a task as the data requirements for this are too significant - the scope of questions can be too wide, especially for the newer domains with no existing wellbeing questions. In addition, the paper discussed a few potential methods for assessing generated questions, arguing that expert evaluation could serve as the more applicable method compared to user studies or automatic evaluation indicators. Lastly, a website was built, showcasing the potential feel and look of the question generation process.

The intersection of AI and wellbeing seems to have enormous potential for both researchers and respondents. In the nearby future, this research could improve through expert interviews or user studies to gain better insight into the usability of transformer-generated questions and how to improve the generation process. For future research, AI-based assessment of answers to open wellbeing questions could be an interesting direction since transformers are shown to provide great results in sentiment detection. This would allow researchers to make use of more open questions without the need to manually extract results from the answers. In addition, this could serve as a small part of an advanced AI-based solution (e.g a wellbeing chatbot) that could pose questions, understand answers and come up with potential solutions - a fully independent AI tool serving a bigger community.

References

[1] A. Brodeur, A. Clark, S. Fleche, and N. Powdthavee, "Covid-19, lockdowns and wellbeing: Evidence from google trends," *Journal of Public Economics*, vol. 193, 01 2020.

[2] W. van der Maden, S. Qui, J. Lomas, and U. Gadiraju, “Using a conversational user interface to assess wellbeing,” *CUI Conference*, 2022.

[3] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, “A systematic review of automatic question generation for educational purposes,” *International Journal of Artificial Intelligence in Education*, vol. 30, 11 2019.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762

[5] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *arXiv:2005.14165 [cs]*, Jul. 2020, arXiv: 2005.14165. [Online]. Available: http://arxiv.org/abs/2005.14165

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[7] K. Elkins and J. Chun, “Can GPT-3 Pass a Writerâs Turing Test?” *Journal of Cultural Analytics*, vol. 5, no. 2, Sep. 2020. [Online]. Available: https://culturalanalytics.org/article/17212-can-gpt-3-pass-a-writer-s-turing-test

[8] R. Thoppilan *et al.*, “LaMDA: Language Models for Dialog Applications,” *arXiv:2201.08239 [cs]*, Feb. 2022, arXiv: 2201.08239. [Online]. Available: http://arxiv.org/abs/2201.08239

[9] “OpenAI API.” [Online]. Available: https://beta.openai.com

[10] A. Romero, “GPT-4 Is Coming Soon. Hereâs What We Know About It,” Apr. 2022. [Online]. Available: https://towardsdatascience.com/gpt-4-is-coming-soon-heres-what-we-know-about-it-64db058cfd45

[11] D. Burns, N. Dagnall, and M. Holt, “Assessing the impact of the covid-19 pandemic on student wellbeing at universities in the united kingdom: A conceptual analysis,” *Frontiers in Education*, vol. 5, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/feduc.2020.582882

[12] “Shaping a brighter world of work: The employer outlook | zurich insurance.” [Online]. Available: https://www.zurich.com/en/knowledge/topics/future-of-work/shaping-a-brighter-world-of-work-the-employer-outlook

[13] J. M. Johnson, D. N. Bristow, and K. C. Schneider, “Did You Not Understand The Question Or Not? An Investigation Of Negatively Worded Questions In Survey Research,” *Journal of Applied Business Research (JABR)*, vol. 20, no. 1, Jan. 2011. [Online]. Available: https://clutejournals.com/index.php/JABR/article/view/2197

[14] J. Winkler, D. Kanouse, and J. Ware, “Controlling for acquiescence response set in scale development,” *Journal of Applied Psychology*, vol. 67, pp. 555–561, 10 1982.

[15] R. Colosi and U. Bureau, “Negatively worded questions cause respondent confusion i,” *Proc Surv Res Methods Sect, Am Stat Assoc, Minneapolis, MN*, 01 2005.

[16] M.-J. Linton, P. Dieppe, and A. Medina Lara, “Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time,” *BMJ Open*, vol. 6, p. e010641, 07 2016.

[17] H. Kunichika, T. Katayama, T. Hirashima, and A. Takeuchi, “Automated question generation methods for intelligent english learning systems and its evaluation,” in *Proc. of ICCE*, 2004.

[18] D. C. Tsai, A. Y. Huang, O. H. Lu, and S. J. Yang, “Automatic Question Generation for Repeated Testing to Improve Student Learning Outcome,” in *2021 International Conference on Advanced Learning Technologies (ICALT)*. Tartu, Estonia: IEEE, Jul. 2021, pp. 339–341. [Online]. Available: https://ieeexplore.ieee.org/document/9499754/

[19] M. Mosbach, M. Andriushchenko, and D. Klakow, “On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines,” 2020. [Online]. Available: https://arxiv.org/abs/2006.04884

[20] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

[21] E. Giaccardi and P. J. Stappers, *Research through Design*, 01 2017, pp. https://www.interaction–design.org/literature/book/the.

[22] “Prompt engineering,” May 2022, page Version ID: 1089715529. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Prompt_engineering&oldid=1089715529

[23] “Notion â One workspace. Every team.” [Online]. Available: https://www.notion.so/product

[24] “Fine-tuning - OpenAI API.” [Online]. Available: https://beta.openai.com/docs/guides/fine-tuning

[25] “From zero to finetuned model in 1 hour with GPT-3. Generate a movie script from any premise!” [Online]. Available: https://www.youtube.com/watch?v=cOz3QJT1zU8

[26] “How to deal with change and uncertainty during COVID-19 - Every Mind Matters,” May 2021. [Online]. Available: https://www.nhs.uk/every-mind-matters/coronavirus/how-to-manage-change-during-covid-19/

[27] Z. Wang, A. Lan, W. Nie, A. Waters, P. Grimaldi, and R. Baraniuk, “Qg-net: a data-driven question generation model for educational content,” 06 2018, pp. 1–10.

[28] J. Baumgartner, N. Ruettgers, A. Hasler, A. Sonderegger, and J. Sauer, “Questionnaire experience and the hybrid System Usability Scale: Using a novel concept to evaluate a new instrument,” *International Journal of Human-Computer Studies*, vol. 147, p. 102575, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1071581920301774

[29] “Personal well-being in the UK - Office for National Statistics.” [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2020tomarch2021

[30] “Coping with loneliness during COVID-19 - Every Mind Matters,” May 2021. [Online]. Available: https://www.nhs.uk/every-mind-matters/coronavirus/coping-with-loneliness-during-coronavirus/

[31] Shreyashankar, “Shreyashankar/gpt3-sandbox: The goal of this project is to enable users to create cool web demos using the newly released openai gpt-3 api with just a few lines of python.” [Online]. Available: https://github.com/shreyashankar/gpt3-sandbox

[32] “Welcome to Flask â Flask Documentation (2.1.x).” [Online]. Available: https://flask.palletsprojects.com/en/2.1.x/

[33] “React â A JavaScript library for building user interfaces.” [Online]. Available: https://reactjs.org/

[34] S. DâAlfonso, “AI in mental health,” *Current Opinion in Psychology*, vol. 36, pp. 112–117, Dec. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352250X2030049X

[35] P. Gooding, “Mapping the rise of digital mental health technologies: Emerging issues for law and society,” *International Journal of Law and Psychiatry*, vol. 67, p. 101498, Nov. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0160252719300950