# DAISI: the Deep Artificial Intelligence System for Interviews

**Hyechan Jun**
Department of Computer Science
Calvin University
Grand Rapids, MI 49546
hyechanjun@gmail.com

**Ha-Ram Koo**
Department of Computer Science
Calvin University
Grand Rapids, MI 49546
haramkoo@gmail.com

**Advait Scaria**
Department of Computer Science
Calvin University
Grand Rapids, MI 49546
ajs244@students.calvin.edu

## Abstract

Natural Language Processing (NLP) has evolved explosively in the past decade as better techniques for training Artificial Intelligence models were discovered and immense amounts of data were accumulated through the internet. We now stand at the tipping point of a potential breakthrough in AI; NLP is poised to create a new generation of virtual assistants, teachers, and perhaps even storytellers. We sought to harness this technology to create an AI Interviewer capable of listening to a story and asking meaningful, appropriate questions. By fine-tuning existing models, creating quantitative analyses of their performance, and refining the models in an iterative manner, we were able to create a prototype of a semi-capable AI Interviewer.

## 1 Introduction

Of the various fields present in AI research, NLP has arguably taken the greatest stride in the past decade. Though Computer Vision, GANs, and other topics have made headway in research circles, NLP has seen the most growth, especially with the introduction of the transformer architecture and the incredible performance of models like BERT, XLNet, or GPT-3 [2] [8]. These models have been proven to score highly on several NLP tasks, including summarization, question-answering, or text classification. GPT-3 in particular has been shown to excel at general language tasks as well, capable of holding a coherent (if not entirely believable) conversation with a human [1]. Not only that, the model is capable of generating creative fiction, poetry, and even computer code, traits that are both incredible and alarming. With this much potential, the logical next question to ask is: where could such technology be used so as to enrich human life without damaging people's livelihoods or otherwise causing harm?

One route that we saw potential for was creating an AI Interviewer. Our initial reasoning was simple: interviewing was a relatively easy task for an NLP system to execute at a basic level, and there was plenty of data in the form of publicly available interview transcripts, podcasts, and so on that could be used as training data. We also reasoned that an interview AI could be utilized as an aid to human interviewers, perhaps generating questions and angles that the human had not considered. This technology could even be used in situations where using a human interviewer was not feasible, such as when a large number of people had to be interviewed simultaneously or an interview had

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

to be conducted in a dangerous location. Our interview AI could also serve as the basis of other NLP applications, such as aiding writers by asking questions about their work in progress. A good interview AI could potentially even lead to AI-based teaching, where the AI asks pertinent questions about academic material. Thus, our goal was to create a prototype of an AI interviewer to see how feasible our idea was and whether it could be expanded in the directions we foresaw. We gave this AI interviewer the working title "DAISI," the Deep Artificial Intelligence System for Interviews.

## 2 Process

Our research process was fairly straightforward: models capable of conducting an interview task already exist, and for our purposes, we just had to find the right data and limitations to fine-tune those existing models on. Figure 1 depicts a simple flowchart of our research process.

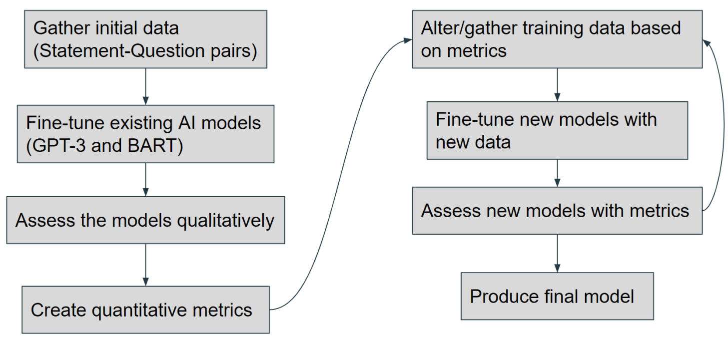


Figure 1: Flowchart of Research Process

### 2.1 Data Wrangling

For our initial dataset, our advisor, Professor Kenneth Arnold, suggested that we utilize the NPR Media Dialog Transcripts compiled by Majumder et al. [7] [6]. We extracted all utterances from this dataset where the guest said something and the host asked the guest a question based on what they said, compiling those utterances into a JSON file of Statement-Question pairs. In essence, this meant extracting all pairs of statements where the guest spoke and the host's statement ended in a question mark. This question-mark data was then separated into a training dataset and an evaluation dataset, and one hundred of the statement-question pairs in the training dataset were randomly selected and set aside to serve as test cases for us to use in qualitative evaluations of each model's performance.

### 2.2 Deciding the Models

The next step to creating our interview AI was choosing the right models to fine-tune. Plenty of models exist in the world today, and we wanted to settle on maybe two or three of them to begin our work. To that end, we first tried to break down what exactly the task of asking interview questions required so that we could utilize models that best fit the task. With help from our advisor, we determined that an interview question-asking task was most like a traditional summarization or translation task, both of which had already been extensively researched and had pre-trained models for. Though we played with the idea of translation using MarianMT, we decided to utilize summarization models for our initial prototypes, shelving the translation idea for later.

The two summarization models we chose were GPT-3 and BART [4]. GPT-3 was a fairly obvious choice; it is one of the most powerful models available today, and because we had access to OpenAI's beta, we saw no reason to not utilize its power. BART was chosen as a counterpart because it has consistently performed well on summarization tasks [3] but utilizes a different architecture with a bidirectional encoder and left-to-right decoder. BART is also an older model trained on far less data, so we were interested to see how that would affect its performance.

### 2.2.1 GPT-3

Fine-tuning GPT-3 was relatively easy thanks to OpenAI's simple API, but unfortunately there was a token limit for fine-tunes and so we were only able to send 20,000 statement-question pairs as training data. We were hoping that GPT-3's extensive capabilities, as demonstrated by its ability in few-shot, one-shot, or zero-shot learning, would be sufficient to overcome the meager training data. We also trained a second GPT-3 model with generic names, courtesy of the SpaCy library. The generic names model was designed to remove potential biases or inclinations in the data based on trigger-words like "Trump" or "Google."

******In December of 2021, we received some good news when OpenAI officially stated that they would lift the limit on tokens for fine-tuning datasets, and so are planning to train new models without the limitations. This paper is a work-in-progress, and this section will be replaced with the update on these larger models.***

### 2.2.2 BART

We trained five different models for BART: one with a training dataset limited to 20,000 lines to match the GPT-3 dataset, one with the full 80,000 lines of training and evaluation data, one with generic names, one with added tokens to represent the length of the expected response, and one with data that extended beyond statement-question pairs. That last model was trained using the full NPR dataset without filtering for when the host asked a question, our aim being to see how the AI's behavior changed when it wasn't trained purely to ask questions.

These models were trained in an iterative fashion, each one aiming to solve a problem or explore a quirk that was evident in the last. The next section, "Qualitative Analysis of Results," will delve further into those problems and quirks.

******Once OpenAI lifts the token limit on GPT-3, we will train more models using their fine-tuning API to reflect the number of models we have for BART.***

## 3 Results

With each model we trained, we created six metrics which we thought were pertinent to asking good interview questions: Appropriateness, Specificity, Repetition, Inclusion of New Concepts, Relevance, and Similarity to the Original Question. We then scored our models' performance on these metrics, initially using a qualitative measuring system where we personally assigned scores for each generated question before attempting to find quantitative measures for each metric. For the following section, the term "statement" will be used to describe the interviewee's statements and the term "question" will be used to describe the interviewer's question in response to the interviewee's statement.

### 3.1 Appropriateness

We defined appropriateness as the measure of how well the question generated by the AI furthered the conversation and set up for further inquiry. For instance, Table 1 depicts a statement and two generated questions—one from BART and one from GPT-3—that have varying levels of appropriateness. GPT-3 was able to form an appropriate question that acknowledged the original interviewee's statement and built off of it to further the conversation. By contrast, BART was able to form a question, but it was quite generic and lacked the ability to push the conversation forward in a meaningful direction.

Quantitative measures for appropriateness proved difficult to find for a variety of reasons. Whether any particular question would continue a conversation effectively was entirely subjective and prone to change depending on the context in which the question was asked and the personalities of the interviewer and interviewee. Though it was relatively easy to spot an inappropriate question, it was more difficult to verbalize why that question was inappropriate because the same question may have been appropriate under different circumstances.

### 3.2 Specificity

Specificity is defined as the measure of how much detail is present in the question, whether that detail is pertinent to the statement or not. A question could have a high specificity score even if it was not at

Table 1: Appropriateness: Example of Qualitative Analysis

| | |
|---|---|
| **Original Statement** | Yes. Yeah, actually, whenever I get together with friends of mine who don't work for newspapers, I realize how much more fun I have than they do. There are a lot more pressures. I have to worry about budgets. I have to worry about the careers of other people. But the reality is I come into the newsroom every morning with about 10 things that I want to get done in terms of coverage. I can influence coverage of everything from how we write about music, to how we write about Supreme Court nominees, to how we cover the most interesting city in America. That's a blast. You can't complain too much about that. |
| **GPT-3 Question** | You just put me into a bit of a dilemma. I'm a reporter, too. How do I have fun, but still have a deadline? |
| **BART Question** | What do you want people to know about you? |

all related to the statement because specificity only measures the amount of detail, not the relevance or appropriateness of those details.

As with appropriateness, finding a quantitative measure for specificity is no easy feat because there is a need for implicit understanding and context in order to determine whether a statement is specific or not. One possible measure we looked into was simply recording the length of the question to determine its specificity because longer responses typically contained more details; however, we decided not to follow through with the idea because there is no real correlation between amount of detail and length of question. Further, a long question could merely be long because of unnecessary information or padding with words, which would confound an attempt to measure specificity quantitatively.

### 3.3 Repetition

Repetition was defined as the measure of overlap of words between the question and the statement. If a question contained many words that were in the statement, it would get a high repetition score; if it contained few words from the statement, it would get a low repetition score.

This statistic was the easiest to turn into a quantitative measure because a scoring system already exists for repeated phrases: ROUGE, the Recall-Oriented Understudy for Gisting Evaluation [5], which measures the overlap of words between two texts. By utilizing ROUGE metrics, we were able to quantify the amount of repetition between the statement and the generated question, allowing us to look for patterns and determine how much repetition typically resulted in a good question.

### 3.4 New Concepts

We defined new concepts as the measure of how much information was present in the question that was not initially in the statement. Sometimes the models would randomly include tangentially related information or even flat-out make things up that were not part of the original statement, necessitating the use of this metric.

We found that GPT-3 tended to insert new concepts into its questions far more than BART, likely due to its larger vocabulary and initial training data. To try and mitigate the number of new concepts being introduced, we used SpaCy to make generic versions of each dataset, removing triggering words like proper nouns that may induce the AI into bringing up new information.

### 3.5 Relevance

Relevance is defined as the measure of how closely related the question is to the statement. It differs from Appropriateness in that appropriateness is determined by the ability of the question to keep the interview going while relevance is more focused on the content of the question and how well it matches the statement.

Finding a quantitative measure for relevance has proven difficult, mainly because relevance is not based on any particular pattern inherent to the English language but instead is a measure that requires implicit knowledge of the topic being discussed and the web of connections between the statement and the question that is not necessarily visible by just examining the words. Our best attempt involves training a model on blank statements and seeing how its questions change as a result, thereby allowing us to compare the blank-statement model with our regular model and determine what factors help create a relevant question.

### 3.6 Similarity to Original Question

This metric is fairly self-explanatory, comparing the AI's output to the original question asked by the interviewer in the training data. Though the original interviewer is by no means an authoritative example of asking good questions, and the questions in the training data were just one of several dozen that could have been asked, we thought it would serve as a good metric for our models to compare against an actual human's response.

## 4 Interactive Console

As part of the research process, we decided to try creating an interactive console through which we could query the AI by giving it statements and see what kind of questions it asked in return. This interactive console proved to be quite enlightening, giving us data to analyze outside of the one hundred test examples we had set aside. Below is a sample of the interactive console's output:

```
Welcome to the BART Interactive Interview AI!
Tell us an interesting story, or type 'END' to stop.
> Hello. My name is Joe Schmoe, and I'm a senior at Calvin
University.

AI: And Joe Schmoe is a senior at Calvin University. How
did he get into the computer science department?
> I'm not in the computer science department; I'm a
Philosophy major.

AI: So you're not in the computer science department;
you're in the philosophy department?
> Yes, I'm actually a double major in Philosophy and
Religion.

AI: You're a double major in philosophy and religion?
> Yes.

AI: What do you mean by that?
> Yes means yes.

AI: What do you mean by that?
> Yes?

AI: What do you mean by that?
> Yes!

AI: What do you mean by that?
> END

Goodbye!
```

As is evident from this short sample, our current model is capable of forming coherent questions, though only to a limited extent. It still suffers from introducing new concepts (the computer science department was never mentioned in the interviewee's statement) and has trouble figuring out what to do about shorter statements such as the simple "Yes." The biggest problem, however, is that the model has no recall ability. It only works off of the last statement the interviewee gave, making it very bad at understanding inter-statement context and thus bad at giving appropriate, relevant answers. This is a large area of improvement that will likely boost the model's scores on all metrics if we can get it right.

## 5 Future Work

Thus far, we have only worked through the first iteration of DAISI, though the next stages should be faster and easier now that we are developing a solid baseline. The most important part of our future work is to continuously iterate through better and better models, gathering new data and shaping it in ways that will make each iteration more useful than the last. To do so, we are considering reducing the number of metrics we use by combining those that are similar to each other. We also plan to utilize GPT-3 more now that the fine-tuning token limit has been lifted. Lastly, we hope to create a better interactive interface for our model, perhaps hosted on a website.

## 6 Conclusion

This project is still in its infancy, with a definite proof of concept that could lead to expansive future work. Though we may not create a perfect AI interviewer in this project, we will create a baseline for further projects like this one, and perhaps create automated measures that can be utilized to quickly train and appraise an AI model in the future.

## Acknowledgements

We would like to thank Professor Kenneth Arnold for acting as our advisor during the course of this project and the Computer Science department at Calvin for providing the resources necessary to see this project through to the end.

## References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. What have we achieved on text summarization? *arXiv preprint arXiv:2010.04529*, 2020.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[5] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[6] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Interview: A large-scale open-source corpus of media dialog. *arXiv preprint arXiv:2004.03090*, 2020.

[7] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, 2020.

[8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.