MInDS: Using Large Language Models to Screen for Depression

Chase Carstensen
Mathematical/Computer Sciences
Worcester Polytechnic Institute
Worcester, MA, USA
wcarstensen@wpi.edu

Nicholas Small
Mathematics/Computer Science
Providence College
Providence, RI, USA
nsmall@friars.providence.edu

Janya Bhaskar
Statistics/Data Science
University of Colorado Boulder
Boulder, CO, USA
jabh4572@colorado.edu

Becks Lopez
Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
rlopez2@wpi.edu

Avantika Shrestha
Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
ashrestha4@wpi.edu

Elke A. Rundensteiner
Computer/Data Sciences
Worcester Polytechnic Institute
Worcester, MA, USA
rundenst@wpi.edu

***Abstract*—Depression is a debilitating, yet underdiagnosed mental illness due to the subjectivity of current screening and time and resource restrictions. Large language models (LLMs) can potentially address these difficulties. Using the Extended Distress Analysis Interview Corpus dataset, containing 105 interview transcripts, we propose MInDS, an automated, modular LLM inferencing pipeline, to optimize depression screening. Our results indicate that LLMs can effectively screen for depression with a 0.8 balanced accuracy. LLM inferencing with a shortened transcript can perform similarly to inferencing with the entire transcript. Our findings may aid the future development of LLMs for depression screening.**

***Index Terms*—LLM Prompting, Classification, Zero-Shot.**

I. Introduction

Depression is a serious mental health condition that has become increasingly common in recent years [1]. Estimates between 2004 and 2014 suggest that roughly 15% of individuals would suffer from depression at some point in their lives [2]. While this was already a large portion of the population, this number started to increase during the COVID-19 pandemic [3], reaching 25% of the population in 2021 [1].

There is a growing concern that it is becoming more difficult for people to seek mental health treatment. Individuals with delayed mental health treatment have worse treatment outcomes than those who receive a more streamlined process [4], and a large portion of depression cases are never diagnosed, which limits treatment options [5]. There is substantial evidence that suggests that treatment for mental health conditions, like depression, help reduce the risk of suicide [6], which further supports a need to improve mental health care.

Three of the most prevalent factors for the under-diagnosis of depression include the subjectivity of the depression screening process [7], a lack of time for both the patient and the clinician [8], and a lack of resources available to the patient [8]. This all leads to a small portion of depressed individuals seeking adequate care [9].

One way to begin to overcome these barriers is through an automated screening process. With a more streamlined process, depression screening could be more consistent and objective by mitigating human error and reducing human intervention. An automated process would therefore eliminate several of the aforementioned barriers to depression screening, thus allowing for more individuals to seek the treatment they need.

Early research into AI screening for depression involved machine learning models, which showed promising results for further investigation [10]. Since then, there has been an increase in research into using large language models (LLMs) to screen for depression, such as ChatGPT 3.5 and 4, Bard/Gemini, Claude, BERT and Llama 2 [11]–[13]. While there is some concern about bias and trustworthiness in LLMs [14], there is a push to explore these models further to determine how to reform mental health treatment [15]. There has also been variation in terms of the types of data used in these models, including passive data from a person's activity, sleep, and social interactions [16], social media posts [17], [18], and interview data [11], [13].

We propose the MInDS system, a modular, versatile, and scalable LLM inferencing pipeline. Our experimentation incorporates two state-of-the-art open source LLMs: Meta's Llama 3 [19] and Google's Gemma 2 [20]. We seek to determine if Llama 3 and Gemma 2 show promise in their ability to perform a task as complex as mental health screening. In addition, we look to determine if we can shorten the interviews used by our MInDS system by asking fewer questions, saving time and resources for all parties involved. Our findings include:

- Gemma 2, with a top balanced accuracy of 0.8, outperforms Llama 3, with a top balanced accuracy of 0.78.
- Gemma 2 performs more consistently across prompts than Llama 3, with Gemma 2's lowest balanced accuracy being 0.72 and Llama 3's being 0.52 across all 104 prompts and hyperparameter settings.

978-1-6654-7345-3/22/$31.00

- With a reduced transcript, Gemma 2 and Llama 3 both achieved a balanced accuracy score within 0.01 of their top balanced accuracy with a complete transcript, indicating that shorter transcripts are promising.

II. Data

A. E-DAIC

In 2014, the University of Southern California developed a dataset called the Extended Distress Analysis Interview Corpus (E-DAIC) [9], which consists of "Wizard of Oz" style clinical interviews between a human-controlled virtual agent named Ellie and a patient [9]. Interviews ranged from 7 to 33 minutes long, with an average length of 16 minutes [9]. We leverage this dataset for use in our research.

In addition to participating in an interview with Ellie, each patient filled out a standardized depression screening survey, the Patient Health Questionnaire (PHQ-8). The PHQ-8 is composed of eight questions regarding a person's mental health. Each question is self-scored between 0 and 3, with 0 meaning they do not identify with the symptoms, and 3 meaning they completely identify with them, with the total score found by summing the patients individual question scores. It is common practice to label anyone with a total score of 10 or above as depressed. The results of the PHQ-8 were used as the ground truth labels for all testing.

The E-DAIC contains no information about the age, race, gender, or any other demographic characteristics of the participants. While this helps to preserve participant anonymity, as a consequence of we cannot confirm that our model is unbiased and performs equally among all demographics.

B. Data Preprocessing

Prior research done by Toto et al. [21] divided the interviews in the E-DAIC into sub-datasets using the responses to the 19 core questions within the transcripts and any follow-up questions. Due to the unique dynamic of each interview, not every patient provided information relating to all 19 categories. They also isolated patient responses. The data used in this paper was the result of these preprocessing steps, where all patient responses had been aggregated and separated by topic.

III. Methodology

A. Large Language Models

Large language models (LLMs) represent a major breakthrough in the fields of machine learning (ML) and natural language processing. LLMs are trained on vast amounts of text data, which allows them to generate human-like responses to input prompts. We used two state-of-the-art open-source LLMs, Meta's Llama 3 and Google's Gemma 2. We accessed these models through the Hugging Face (HF) Hub. We deployed open-source LLMs because these models can be run locally, allowing us to keep confidential patient interview data secured on our own servers.

1) Llama 3: In April 2024, Meta unveiled, at the time, their largest and most capable open-source LLMs with the release of the Llama 3 family of models [19]. The models came in two different sizes: 8 billion parameters, and 70 billion parameters. Meta also released versions of each model fine-tuned to accurately follow instructions. We focused on the 8 billion parameter instruction tuned model, Llama 3 8B Instruct.

2) Gemma 2: Similarly, in June 2024, Google released their Gemma family of open-source LLMs [20]. These models also came in two sizes, 9 billion parameters and 27 billion parameters, along with instruction-tuned variants. We used the 9 billion parameter instruction tuned model, Gemma 2 9B It.

B. Prompt Engineering

1) Creating Prompts: We chose to break the prompt into three components: the identity, the job, and the output. This is based on the RISE prompting framework: Role, Input, Steps, and Expectation.

The identity is where the user tells the model what role to assume when conducting the assigned task, such as "therapist", "counselor", and "psychologist".

The job informs the model of the assigned task. This portion of the prompt comprises both the Input and Steps sections. During our study, we told the model to analyze an interview transcript and then classify the patient as being depressed or not depressed. In order to increase prompt variability, in some of the prompts we provided the LLM with detailed criteria for labeling someone with depression, and for others we simply told the LLM to use what it already knows about depression.

Finally, the model is instructed on the desired output format. All prompts in our study were given the same output format instruction: "Respond with only 'depressed' or 'not depressed'." This allows the output to be easily encoded later during post-processing.

We compiled a list of 13 unique identities, 8 unique jobs, and a single output format. By putting together each possible combination of identity, job, and output, we generated 104 prompts.

2) Transcript Generation: As mentioned in subsection II-B, our dataset is comprised of isolated patient responses to a variety of interview questions. When compiling our interviews, short of re-transcribing all interview audio recordings, we did not have access to the exact questions the virtual agent asked each patient. We used a list of question descriptions [21], to form our pseudo-transcripts. Table I contains the question identifiers and generic questions used in this study.

C. The MInDS System

We now introduce our proposed pipeline, Modular Inference for Depression Screening, or MInDS. With a large number of candidate prompts, it is important to design a system with which we can easily and efficiently test them. The MInDS system greatly increases prompt engineering and testing speeds.

TABLE I: Question identifiers and generic questions used for pseudo-transcripts

Identifier	Generic Question
'doing_today'	How are you doing today?
'argued_someone'	When was the last time you argued with someone and what was it about?
'advice_yourself'	What advice would you give yourself ten to twenty years ago?
'proud_life'	What are you most proud of in life?
'controlling_temper'	How are you at controlling your temper?
'last_happy'	When was the last time you felt really happy?
'easy_sleep'	How easy is it for you to get a good night's sleep?
'friend_describe'	How would your best friend describe you?
'dream_job'	What is your dream job?
'study_school'	What did you study in school?
'anything_regret'	Is there anything you regret?
'travel_lot'	Do you travel a lot?
'diagnosed_depression'	Have you ever been diagnosed with depression?
'feeling_lately'	How have you been feeling lately?
'diagnosed_p_t_s_d'	Have you ever been diagnosed with PTSD?
'do_relax'	What do you do to relax?
'positive_influence'	Tell me about a positive influence in your life.
'dont_l_a'	What are some things you don't like about LA?
'move_l_a'	Why did you move to LA?

1) General Outline: MInDS is comprised of three distinct steps, the first of which is the prompt preprocessing phase. In this phase the prompts generated during subsection III-B are correctly formatted. Additionally, we compiled the interview transcripts for all 105 patients using a question and answer format. Here, we can instruct MInDS on what questions should be included in the interview transcript. For example, if the pipeline is instructed to include the 'doing_today' and 'feeling_lately' questions, then the resulting transcript would only include those corresponding questions. Figure 1 shows an example of a completed prompt and transcript.

```
You are a therapist tasked with screening patients
for depression. Read the following interview
transcript and predict if the patient has
depression. Respond with only 'depressed' or 'not
depressed'.
Q: How are you doing today?
A: PATIENT RESPONSE
Q: How have you been feeling lately?
A: PATIENT RESPONSE
```

Fig. 1: Example prompt

The second step of MInDS is the model inference phase. The user can instruct MInDS to use any LLM on the HuggingFace Model Hub with text generation capabilities, given that the user's access token has read permissions to the model's repository. We ran all LLMs at full precision on the Turing Cluster, the main high performance computing cluster of the Academic & Research Computing (ARC) group at WPI. This allowed us to spread LLM inference across multiple GPUs which greatly decreased the compute time. For each prompt, the LLM performs 105 inferences, one for each patient transcript.

The final phase of our pipeline is the processing of the model output by encoding the output and calculating metrics. The user can instruct MInDS to encode the output and calculate metrics in any way desired. MInDS has built in support for accuracy, precision, recall, specificity, F1, and balanced accuracy. It also has support for running the complete system any given number of times and averaging the outputs, then creating visualizations of the top performing prompts.

A full run of the system included each of these three steps in order. In all, the prompts and transcripts were formatted and assembled, the given LLM performed inferences on each prompt with each interview transcript, and finally the output was processed.

2) Versatility, Scalability, and Automation: The MInDS system is controlled via a command line interface (CLI). All input, including prompts, interview data, and LLM of choice can be set using the MInDS CLI. Additionally, MInDS contains a basic scheduler to distribute jobs across all available CUDA enabled devices. Through our access to WPI ARC's Turing Cluster, we had access to NVIDIA H100 GPUs which were used for all LLM inferencing. Further, we often ran the full system multiple times and averaged the output to generate a robust result. It was common for our use of the MInDS system to require an LLM to perform over 100,000 inferences. Our estimates are that throughout all experiments, we inferenced various LLMs over 1 million times. This amount of experimentation would not have been possible without a flexible, efficient and streamlined prompting system.

3) Prompt Formatting Differences Between LLMs: Both Llama 3 8B Instruct and Gemma 2 9B It have their own chat templates which can be used to optimize prompt formats for use with the LLM. The key difference between the prompt formats for the two LLMs used is their support of a system prompt. Llama 3 8B Instruct supports the use of a system prompt to guide LLM inferencing, but Gemma 2 9B It does not. MInDS has the functionality to automatically use an LLM's default chat template to format prompts with the appropriate system prompt usage.

D. Question Selection

In order to reduce both the patient interview time and compute time, we implemented a greedy question selection algorithm that would shorten the transcripts by selecting a subset of questions from the entire interview transcript and determining which produced the highest balanced accuracy using the MInDS system. Algorithm 1 details our question selection implementation.

IV. Experimental Results

A. Experimental Setup

Our experimental study using MInDS includes evaluation (A) on complete transcripts comparing Llama 3 and Gemma 2 across different prompting styles, and (B) on subsets of transcripts composed of few questions.

We use the balanced accuracy (BA) metric to evaluate the effectiveness of our experiments defined as the mean of sensitivity (true positive rate) and specificity (true negative rate). BA is particularly robust when used with imbalanced datasets, such as the E-DAIC dataset.

Algorithm 1 Greedy Question Selection Algorithm

Require: $k \geq 1$
  qs_left ← all question identifiers
  qs_used ← empty list
  while Length qs_used $< k$ **do**
    results ← empty list
    for question in qs_left **do**
      results ← results + MInDS(qs_used + question)
    end for
    top_question ← question with the highest balanced accuracy in results
    qs_used ← qs_used + top_question
    qs_left ← qs_left - top_question
  end while
  return qs_used

We found that Gemma 2 produced more consistent results compared to Llama 3. Specifically, whenever Gemma 2 performed inference on a particular prompt, it would return the same predictions. Whereas with Llama 3, there were variations to the output. Thus we only ran each experiment with Gemma 2 once, while we ran Llama 3 experiments between 5 to 10 times and averaged the result metrics. As previously mentioned, this functionality was built into MInDS.

B. Comparing LLMs across Different Prompting Styles

As can be seen in Table I, Gemma 2 performed slightly better than Llama 3, with Gemma 2's top BA score being 0.8 using the prompt in Figure 3. In contrast, the top BA achieved by Llama 3 was 0.78 using the prompt in Figure 2, with the job in red, the identity in blue, and the output format in black.

```
You are a therapist who is an expert at identifying
patients with depression. Your task is to label
patients as depressed or not depressed with very
high accuracy. To label patients as 'depressed'
they should display either a loss of interest in
daily activities or a depressed mood which
interferes with their quality of life in the
following interview transcript. Label all patients
as 'not depressed' who do not exhibit any of the
previously mentioned symptoms. Respond with only
'depressed' or 'not depressed'.
```

Fig. 2: Best-performing Llama 3 prompt

When comparing the top prompts for the different LLMs, as in Figure 2 and Figure 3, the individual prompts that performed well are different between the two models, as with the prompts that performed poorly.

Gemma 2 performed more consistently across prompts than Llama 3 did, as shown in Figure 4, with the standard deviation of Gemma 2 across all prompts being 0.0173 compared to 0.08 for Llama 3. To generate this visualization, the results from Gemma 2 and Llama 3 were sorted by BA. For each model, we then graphed the BA of the prompts with the maximum BA, the median BA, and the minimum BA. Overall, Gemma

```
You are a psychiatrist who is trying to accurately
diagnose your patients with depression. You are
very good, and want only what is best for your
patients. You are relieved and very happy when a
patient is correctly diagnosed. Classify each
patient into one of two categories. If the patient
appears to suffer from depression, respond with
'depressed'. If the patient does not appear to
suffer from depression, respond with 'not
depressed'. Respond with only 'depressed' or 'not
depressed'.
```

Fig. 3: Best-performing Gemma 2 prompt

2 outperformed Llama 3 across the board. This shows that the choice of model also has a large impact on the result of the MInDS system.

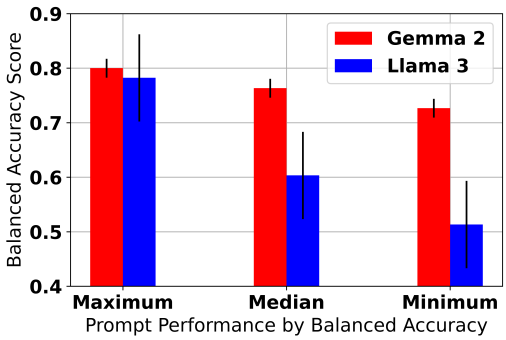


Fig. 4: Model comparison by prompt performance

C. Comparing LLM Models across Different Transcript Sizes

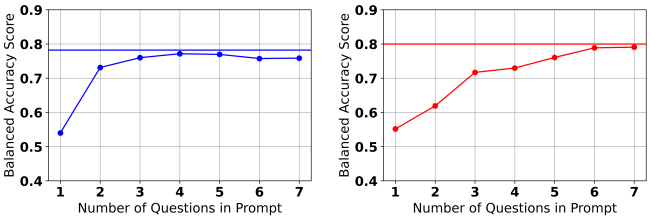
The greedy question selection algorithm described in Algorithm 1 runs the MInDS system several times. To improve the computing time of the algorithm, we compiled a list of the 10 highest-performing prompts for each LLM based on the results of a full run of the MInDS system. It was with this list that the question selection algorithm was run.

As mentioned earlier, the highest BA achieved with Llama 3 was 0.78. After running our question selection algorithm, we found that the best-performing transcript subset consisted of just 4 questions, and achieved a BA of 0.77. BA fell slightly after that, which can be seen in Figure 5a. The questions, listed by identifier, that were selected by our algorithm were: 'diagnosed_depression', 'friend_describe', 'controlling_temper', and 'anything_regret'.

The highest BA achieved with Gemma 2 was 0.8. Gemma 2's best transcript subset, after running the question selection algorithm, consisted of 7 questions and had a BA of 0.79, as can be seen in Figure 5b. The questions were: 'last_happy', 'feeling_lately', 'diagnosed_p_t_s_d', 'friend_describe', 'diagnosed_depression', 'controlling_temper', 'dream_job'.

D. Discussion

While LLMs performing inference transcripts solicited via virtual agents could be incorporated into healthcare to expedite the broad screening for mental health conditions, subsequent



(a) Llama 3 (b) Gemma 2

Fig. 5: BA for Different Transcript Sizes Composed of Different Numbers of Questions-Response Pairs.

clinician time for a more in-depth diagnosis will remain essential. Also, great care must be taken that such data is protected securely from malicious agents, i.e., transcripts should not be collected without the knowledge of the creator nor used for discriminatory or marketing purposes. This is an important ethics challenge, and our use of LLMs instead of other ML techniques does not increase the risk of patient data misuse. We address the confidentiality concern in part here by the utilization of open-source LLMs that are run locally on protected servers to allow us to keep data secured instead of sending the patient data to proprietary servers such as OpenAI's GPT models.

V. Conclusion

In this work, we explored the viability of using Large Language Models (LLMs) for screening for mental health illnesses such as depression on interview transcript data. For this, we constructed the MInDS system, an end-to-end, modular LLM inferencing pipeline, designed for ease of experimentation. We utilized MInDS to explore the effectiveness of alternate prompting styles for optimizing depression screening and evaluated the selection of different question-topic subsets in place of the full interview transcript in order to shorten interview length. Our results demonstrate that LLMs are effective in depression screening. Gemma 2 performed with a top BA score of 0.8, which was consistently better than Llama 3 for the diverse styles of prompting we explored. When using a four to six question-topic subsets of the transcript, we saw similar performance to screening with the entire transcript for Gemma 2 and Llama 3, respectively. In summary, our findings indicate the promise for leveraging modern LLM-based systems for depression screening,and thus encourage future exploration into addressing this important mental health challenge.

Acknowledgments

Thank you for funding from the NSF REU Site No. 2349370, and for computational resources from WPI ARC.

References

[1] J. Bueno-Notivol, P. Gracia-García et al., "Prevalence of depression during the covid-19 outbreak: A meta-analysis of community-based studies," International journal of clinical and health psychology, vol. 21, no. 1, p. 100196, 2021.

[2] G. Y. Lim, W. W. Tam et al., "Prevalence of depression in the community from 30 countries between 1994 and 2014," Scientific reports, vol. 8, no. 1, p. 2861, 2018.

[3] R. Mojtabai, M. Olfson, and B. Han, "National trends in the prevalence and treatment of depression in adolescents and young adults," Pediatrics, vol. 138, no. 6, 2016.

[4] A. Reichert and R. Jacobs, "The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in england," Health Economics, vol. 27, no. 11, pp. 1772–1787, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3800

[5] G. S. Malhi and J. J. Mann, "Depression," The Lancet, vol. 392, no. 10161, pp. 2299–2312, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673618319482

[6] J. M. Bertolote and A. Fleischmann, "Suicide and psychiatric diagnosis: a worldwide perspective." World psychiatry : official journal of the World Psychiatric Association, vol. 1 3, pp. 181–5, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:29519086

[7] K. Kroenke, T. W. Strine et al., "The phq-8 as a measure of current depression in the general population," Journal of affective disorders, vol. 114, no. 1-3, pp. 163–173, 2009.

[8] E. M. Colligan, C. Cross-Barnet et al., "Barriers and facilitators to depression screening in older adults: a qualitative study," Aging & mental health, vol. 24, no. 2, pp. 341–348, 2020.

[9] J. Gratch, R. Artstein et al., "The distress analysis interview corpus of human and computer interviews." in LREC. Reykjavik, 2014, pp. 3123–3128.

[10] Y. Terhorst, L. B. Sander et al., "Optimizing the predictive power of depression screenings using machine learning," DIGITAL HEALTH, vol. 9, p. 20552076231194939, 2023. [Online]. Available: https://doi.org/10.1177/20552076231194939

[11] J. Ohse, B. Hadžić et al., "Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection," Computer Speech Language, vol. 88, p. 101663, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230824000469

[12] Z. Elyoseph, I. Levkovich, and S. Shinan-Altman, "Assessing prognosis in depression: comparing perspectives of ai models, mental health professionals and the general public," Family Medicine and Community Health, vol. 12, no. Suppl 1, 2024. [Online]. Available: https://fmch.bmj.com/content/12/Suppl_1/e002583

[13] M. Sadeghi, B. Egger et al., "Exploring the capabilities of a language model-only approach for depression detection in text data," in 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2023, pp. 1–5.

[14] A. Ferrario, J. Sedlakova, and M. Trachsel, "The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: A critical analysis," JMIR Ment Health, vol. 11, p. e56569, Jul 2024. [Online]. Available: https://mental.jmir.org/2024/1/e56569

[15] M. Omar, S. Soffer et al., "Applications of large language models in psychiatry: a systematic review," Frontiers in Psychiatry, vol. 15, 2024. [Online]. Available: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2024.1422807

[16] Z. Englhardt, C. Ma et al., "From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 2, may 2024. [Online]. Available: https://doi.org/10.1145/3659604

[17] N. Farruque, R. Goebel et al., "Depression symptoms modelling from social media text: an llm driven semi-supervised learning approach," Language Resources and Evaluation, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268950421

[18] X. Xu, B. Yao et al., "Mental-llm: Leveraging large language models for mental health prediction via online text data," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 1, mar 2024. [Online]. Available: https://doi.org/10.1145/3643540

[19] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[20] G. Team, "Gemma," 2024. [Online]. Available: https://www.kaggle.com/m/3301

[21] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, ser. CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 4145–4154. [Online]. Available: https://doi.org/10.1145/3459637.3481895