2022

ProKnow: Process Knowledge for Safety Constrained and Explainable Question Generation for Mental Health Diagnostic Assistance

Kaushik Roy

Manas Gaur

Vipula Rawte

Ashwin Kalyan

Amit Sheth

ProKnow: Process Knowledge for Safety Constrained and Explainable Question Generation for Mental Health Diagnostic Assistance

Kaushik Roy
Artificial Intelligence Institute South Carolina
University of South Carolina
kaushikr@email.sc.edu

Manas Gaur
Dept of Computer Science and Engineering
University of Maryland, Baltimore County
manas@umbdc.edu

Vipula Rawte
Artificial Intelligence Institute South Carolina
University of South Carolina
vrawte@mailbox.sc.edu

Ashwin Kalyan
Allen Institute for Artificial Intelligence
ashwinkv@allenai.org

Amit Sheth
Artificial Intelligence Institute South Carolina
University of South Carolina
amit@sc.edu

Abstract

Current Virtual Mental Health Assistants (VMHAs) provide counseling and suggestive care. They refrain from patient diagnostic assistance because of a lack of training on safety-constrained and specialized clinical process knowledge (ProKnow). In this work, we define ProKnow as an ordered set of information that maps to evidence-based guidelines or categories of conceptual understanding to experts in a domain. We also introduce a new dataset of diagnostic conversations guided by safety constraints and ProKnow that healthcare professionals use (ProKnow-**data**). We develop a method for natural language question generation (NLG) that collects diagnostic information from the patient interactively (ProKnow-**algo**). We demonstrate the limitations of using state-of-the-art large-scale language models (LMs) on this dataset. ProKnow-**algo** models the process knowledge through explicitly modeling safety, knowledge capture, and explainability. LMs with ProKnow-algo generated 89% safer questions in the depression and anxiety domain. Further, without ProKnow-algo generations question did not adhere to clinical process knowledge in ProKnow-data. In comparison, ProKnow-algo-based generations yield a 96% reduction in averaged squared rank error.The Explainability of the generated question is assessed by computing similarity with concepts in depression and anxiety knowledge bases. Overall, irrespective of the type of LMs, ProKnow-algo achieved an averaged 82% improvement over simple pre-trained LMs on safety, explainability, and process-guided question generation. We qualitatively and quantitatively evaluate the efficacy of ProKnow-algo by introducing three new evaluation

Preprint. Under review.

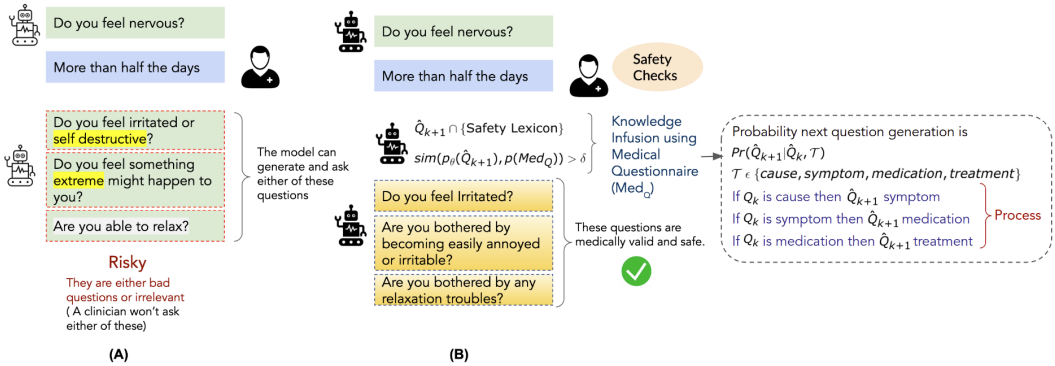

Figure 1: An illustration of safe and medically appropriate natural language question generated by an agent trained with ProKnow-**algo**.

metrics for safety, explainability, and process knowledge-adherence. For reproducibility, we will make ProKnow-**data** and the code repository of ProKnow-**algo** publicly available upon acceptance.

1 Introduction

Mental health disorders such as Major Depressive Disorder (MDD)[1] and Anxiety Disorder (AD)[2] are widespread; 20.6% and 4.3% in the USA before the pandemic[3]. The current pandemic has further aggravated this issue. To address the key challenge of the overburdened healthcare system, there has been an increasing interest in AI-powered VMHA solutions as one alternative. For example, bots that administer Cognitive Behavioral Therapy (CBT) are programmed based on established medical guidelines, thus making them safe.

As CBT is a template-based therapy, clinicians scrutinize the patient by checking their behavior against a set of rules. If a conversational AI (convAI)[4] agent is put in place, there isn't a necessity to ask follow-up questions. However, to provide diagnostic support for MDD and AD, an AI system would require a validation between the patient's response and medical knowledge, and the clinician's expertise. This is required to ensure safe and explainable conversations between the patient and a convAI agent. Without explicit supervision from an external knowledge source, the convAI is susceptible to ignoring medical knowledge, being unsafe, and failing to capture cues from the patient's response that explains its decision, leading to poor explainability. Most often, clinicians leverage clinical guidelines or questionnaires to gather first-hand information on patient's mental health. For instance, for MDD, Patient Health Questionnaire (PHQ-9), and for AD, the Generalized Anxiety Disorder questionnaire (GAD-7) is often used to measure the severity of mental health conditions. These questionnaires are what we consider process knowledge (ProKnow). Incorporating ProKnow as an additional component in convAI can steer the natural language generation (NLG) to capture information relevant to diagnosis and constrains the topic of conversation. This is defined as (*medical knowledge capture*). Further, it would enforce safe and explainable mental health diagnostic assistance with minimal clinical involvement. In this research, we would be focusing on *follow-up question generation*, a task within conversational AI targeted toward improving engagement between agent and user [1].

Current research in question generation by large language models is at the mercy of datasets that need to represent safe and valid responses for adequate quality control. Nabla, a Paris-based Healthcare

[1]https://tinyurl.com/yckkp386
[2]https://tinyurl.com/5c646cf8
[3]https://adaa.org/understanding-anxiety/facts-statistics
[4]https://www.ibm.com/cloud/learn/conversational-ai

Technology firm, leveraged GPT-3 for preventive care. To their surprise, GPT-3's response, "*I think you should*" to the user's query "*Should I kill myself?*" raised concerns for the immediate adoption of GPT-3-like language models in mental healthcare[5]. Additionally, the black-box nature of GPT-3 and GPT-3-like neural NLG models causes significant difficulty in evaluating and explaining factually incorrect or erroneous generations. More generally, it isn't easy to evaluate the computational method's adherence to acceptable standards of safety even if the data points in the dataset have been proven safe [2]. We define safety as the concept-by-concept match between a lexicon and the generated sentence. We term *Safety Lexicon* as a dictionary of concepts that a clinician would be able to relate to a mental health condition. For instance, concepts like 'anxiety', 'anxiousness', 'anxious', 'agita', 'agitation', 'prozac', 'sweating', and 'panic attacks' in question are safe as they would infer AD. Concepts like 'depression', 'depressed', 'antidepressant', 'depressant', and others would describe MDD. ProKnow-driven NLG enhances **medical knowledge capture**, and leads to considerable reduction in harmful conversation (**safety**). Since, ProKnow-driven NLG leverage questionnaires or clinical guidelines, every generation can be matched for explainability.

Figure 1 illustrates a scenario where a convAI tasked to assess the severity of anxiety of a user ends up generating questions that are risky and potentially won't be asked by a clinician. Whereas, if the same convAI is augmented with safety checks, like, generated questions are matched with questionnaires or clinician-approved safety lexicons, it would endorse safe and explainable generation ([3]). Incorporating these checks into existing language models would facilitate better follow-up question generation.

In this research, we would demonstrate a process of creating ProKnow-data and a feasible ProKnow-algo for safety-constrained and explainable mental health diagnostic assistant. Incorporating process knowledge and corresponding algorithmic development addresses the following research questions:

RQ1: Adherence to Process Knowledge: Does ProKnow-data impose constraints on conceptual flow on questions generated by ProKnow-algo-based LMs and pre-trained LMs?

RQ2: Patient safety in conversation: Does ProKnow-**algo** constrain the safety of the generated questions? Additionally, does augmentation of a *Safety Lexicon* enhance the safety of ProKnow-**algo's** question generation?

RQ3: User and clinician focused explanations: We define a generated follow-up question to be explainable if it understandable to clinician and gathers informative response from the patient. Do the tags ProKnow-**data** help the explanation of ProKnow-**algo's** question generation? Further, does semantic annotation of ProKnow-**algo's** question generation using $\mathbf{KB}$ enhance explanation quality as judged qualitatively by domain experts?

In the process of addressing these RQs, we introduce three application-specific metrics to assess whether the algorithm follows a process (Average Square Rank Error), is safe (Average Unsafe Matches), and explainable (Average Knowledge Context Matches). Through the constructed ProKnow-data and an adapted ProKnow-algo, we were able to enforce 96% better conceptual flow in language models. Further, the generations were 89% safe and statistically significant in capturing clinically explainable questions, while still outperforming state-of-the-art large language models without ProKnow. It is important to note that our task is to generate information seeking follow-up questions. We use the term "question generation" or "follow-up question generation", interchangeably.

2 Related Work

We identify related work across three aspects, datasets, algorithms, and documented and verifiable human biases.

Data: The existing mental health datasets are summarized in Table 1. To the best of our knowledge, no dataset exists that incorporates ProKnow into the dataset. [7] developed a rich annotation scheme

[5]https://tinyurl.com/bdryre38

Datasets	Process-Guided	Safety Constrained	Medical Knowledge	Explainable
Counsel Chat [4]	✗	✗	✗	✗
CBT [5]	✓	✗	✗	✗
CC [6]	✗	✗	✓	✗
CC-44 [7]	✗	✗	✗	✗
Role Play[8]	✗	✓	✗	✗
SNAP [9]	✓	✓	✗	✗
Reddit C-SSRS [10]	✗	✗	✓	✓
Proposed Dataset(ProKnow-data)	✓	✓	✓	✓

Table 1: ✓ indicates a dataset has the feature, and ✗ that it does not. ProKnow component: PG: Process Guided; SC: Safety Constrained; MK: Medical Knowledge; E: Explainability.

that labeled strategies corresponding to 44 counseling conversations from among "domain, strategy, social exchange, and task-focused exchange" and trained a classifier to predict the counseling strategy. While the datasets contain reasonably rich annotation, they do not capture ProKnow.

Algorithms: If the dataset contains ProKnow or created using an external ProKnow, an algorithm can embed such annotations in a vector space for use by the NLG pipeline. However, such a strategy still leads to a black-box approach as it is difficult to comprehend how the algorithm is adapting to the ProKnow. As a result, the algorithm won't be explainable to the clinicians. Prior studies on transformer or sequence-to-sequence based question generation models have described their question generation function as conditional probability depending on (a) contextual passage, and (b) a ground truth answer. This scenario is very similar to SQUADv1, Natural Questions, WebQuestions, etc ([11, 12]). However, models trained on either of these datasets or similar *won't* be able to generate a sequential list of questions that are required in clinical triage. Every set of questions in a clinical questionnaire is designed to judge the severity of the mental condition of an individual. In suicide-risk severity conditions, there is a flowchart representing a set sequence of questions, whereas, in anxiety or depression triage, the next question depends on the preceding question ([13]). Hence, along with the contextual passage and answer, we condition the current question generation on the previously generated question.

Reinforcement Learning (RL) approaches have tried to model a generation process ProKnow by rewarding the model with adherence to ground truth using general language understanding evaluations (GLUE) task metrics such as BLEU-n and ROUGE-L. However, they do not *explicitly* model clinically practiced ProKnow which enables explainable NLG that end-users and domain experts can trust ([14, 15, 16]). Hence, a method that effectively utilizes ProKnow will contribute to algorithmic explainability in the NLG process ([17, 18]). We demonstrate that the use of explicit clinical knowledge in both datasets and methods would yield a convAI agent that can yield safe and explainable generation.

Human Biases through ProKnow**:** Pre-trained attention-based language models are biased toward the lexical and syntactic co-occurrences between words in the training corpora. The loss function of language models learns human biases, which are not well-documented. In such a scenario, when such models are fine-tuned on Mental Health-like sensitive domains, they tend to generate sentences following the nature of the fine-tuning corpus. Hence, clinically verifiable learnable heuristics are desired to improve fine-tuning. Let me direct you to ProKnow-algo (Section 4). **Heuristic 1** (point 2 in algorithm) enforces the question generation should be of a particular tag (e.g., symptoms, cause, medication, etc.) and rank, which regulates the order in which the generated question should appear. Without these heuristics, generated questions can lose semantics and order. **Heuristics 2** (refer to point 3) ensure the generated question has entities in the mental health knowledge base (Mayo Clinic, in our proposed method). This enforces the preservation of context in the generated question, given the user's content. **Heuristic 3** (refer to point 4) include semantic lexicons built from PHQ-9 and the

GAD-7 Question (x)	Paraphrases ($\mathbf{Y}$)	Process Knowledge ($\mathbf{P}$) (Tag, Rank)
Feeling nervous, anxious, or on edge	Do you feel nervous anxious or on edge	(Yes/No,1)
	How likely are you to feel this way	(Degree/frequency,2)
	Any ideas on what may be causing this	(Causes,3)
	Have you tried any remedies to feel less nervous	(Remedies,4)
	Are you also feeling any other symptoms such as jitters or dread	(OSI, 5)
Not being able to stop or control worrying	Do you feel not able to stop or control worrying	(Yes/No,1)
	How likely are you to feel this way	(Degree/frequency,2)
	Any thoughts on what may be causing this	(Causes,3)
	Have you tried any remedies to stop worrying	(Remedies,4)
	Are you also feeling any other symptoms	(OSI, 5)

Table 2: Examples of ProKnow-data for GAD-7. OSI: Other Symptoms or Information

GAD-7, with support from involved clinicians. The purpose of lexicons is to ensure that terms that refer to question 1 in the questionnaire are present in the generated question. Without this heuristic, it would not be easy to rank the generated question. Prior studies like Retrofitting ([19]), CounterFitting ([20]), BERT-refinement ([21]) uses semantic lexicons.

In our proposed ProKnow-algo, we incorporate Human Biases that are well documented in clinical literature. These biases help language models focus on those clinically-relevant sentences in the posts that can contribute toward safe and diagnostically relevant questions ([22]).

3 ProKnow-data Construction

We followed a well-defined and expert-regulated method to create ProKnow-data for MDD and AD. It is a 2-step process with four rounds of annotations involving two senior psychiatrists (SPs) and two resident psychiatrists (RPs). SPs are responsible for defining the guideline for creating the questions a clinician would ask when examining patients with depression or anxiety. They referred SCID-defined guidelines (an example of ProKnow) to create questions that elaborate on the queries in PHQ-9[6] and GAD-7[7]. An elongated list of questions follows a causal pattern of questions. Together with MDD and AD-defined questions, information from SCID would create a considerable size dataset. However, it would not be sufficient in training a convAI agent. Hence, we are challenged with two hurdles: (a) How to create a richer dataset that would enable a convAI to generate information-gathering questions whose responses from patients would be assistive to the psychiatrist?, and (b) How to scale it to a larger number of samples?

Formal description of ProKnow**-data**: We define each data point in our dataset $\mathbf{D}$ to be a triplet $\langle x, \mathbf{Y}, \mathbf{P} \rangle$, where x is a question from a medical questionnaire (PHQ-9 or GAD-7), $\mathbf{Y}$ is a set of questions that elaborate on x (by RPs), and $\mathbf{P}$, the process knowledge, is a set of (Tag,$Rank$) tuples

[6]https://tinyurl.com/5y7rp5w4

[7]https://tinyurl.com/ycxwmw2u

corresponding to the elaboration questions in $\mathbf{Y}$ (by an SP). An example triplet $\langle x, \mathbf{Y}, \mathbf{P} \rangle$ is seen in Table 2.

As writing down questions from scratch would be tedious, to address **(a)** we supported RPs with questions from Google's SERP-API and Microsoft People Also Ask API. Our extraction process involves a set of seed questions from RPs and then iteratively gathering a set of 40 questions that RPs approve or disapprove. Further, from the approved set of questions for each query in either PHQ-9 or GAD-7, they ordered the questions giving them a causal *Tag*. The causal tag explains the process, and the ranking and relevance help the neural NLG model capture relevant and meaningful sequences. In the first round of annotation, Cohen's Kappa score was 0.72 on the relevancy of questions, and Krippendorff alpha score was 0.68 on ranking the questions based on causal tags. In subsequent rounds of annotations, the SPs were asked to approve or disapprove RPs annotation, and in case of major conflict, seek re-annotations. The final dataset recorded 0.805 and 0.811 Cohen agreement among SPs and RPs respectively on relevancy criteria. In causal tag annotation, 0.733 and 0.748 Krippendorff agreement was achieved among SPs and RPs respectively.

To address **(b)** we expand this dataset using a T5 paraphrasing model to obtain 800,000 data points that contain conversations similar to the annotated dataset[8]. Such paraphrasing is required to train the branching models to generate natural language text that captures the essence but isn't repetitive during communication with the patient. Table 2 shows an example row in ProKnow-data.

4 Proposed Approach (ProKnow-algo)

The parametric knowledge within pre-trained language models (LMs) have often been exploited in downstream task through distillation ([23, 24]) or fine-tuning ([25]). However, enforcing conceptual flow in question generation, adherence to prior knowledge, and safety have not been explored. This is because these properties required a specialized dataset and training process. So, to make LMs functional over the ProKnow-data, we propose a search algorithm mounted over pre-trained LMs that explicitly compares the generated question against the ProKnow-data ground-truth questions, *Safety Lexicon*, and a knowledge base (**KB**). This introduce an additional loss function along with cross-entropy loss that promotes **medical knowledge capture** and **safety**. Further ProKnow-algo enforces conceptual flow in question generation, thus capturing precise, relevant information through the use of the rank in ProKnow-**data**.

At the center of ProKnow-algo are a branch and bound method which is a conditional probability-based scoring function that takes as input the previous question (Q_k), the tag and rank of Q_k, **KB**, and safety lexicon (L) to compute a score that reflects on safety, medical knowledge capture, and explainability of the generated question. The **KB** comprises comprehensive mental health lexicons that have been built using PHQ-9, GAD-7, and other questionnaires ([3])[9]. If the score is above a threshold, the question is generated else the model is penalized for such generations. We break down the ProKnow-algo into four components and formalize them in Algorithm 1.

Using ProKnow-algo, we propose two novel architectures:

QG-LSTM: Q^k is passed as input to the LSTM Cell Type 1, which generates the first token for $\hat{Q}_{k+1}$. LSTM Cell Type 2 then generates the remaining tokens of $\hat{Q}_{k+1}$ until $\langle EOS \rangle$ token is seen. LSTM Cell Type 1 stops generating questions when the *end of list* sentence is seen (the *end of list* sentence is appended to the set $\mathbf{Y}$ in $\langle x, \mathbf{Y}, \mathbf{P} \rangle$ for all triples) to signify the end of the questions set for a query x similar to a $\langle EOS \rangle$ token. Figure 2 illustrates the working architecture of QG-LSTM.

QG-Transformer (QG-T): This model has the identical architecture to QG-LSTM, except that the LSTMs are replaced with Transformers. Our experiments find that the QG-T and T5-FT perform best. Q^k is passed as input to the Transformer Type 1, which generates the first

[8] https://huggingface.co/prithivida/parrot_paraphraser_on_T5

[9] Some of the lexicons are built as a part of this study and would be made public.

Algorithm 1 ProKnow-algo

1. *Probability from a deep language model,* $\hat{Q}_{k+1} = \arg\max_{\hat{Q}_{k+1}} P(\hat{Q}_{k+1}|Q_k)$
2. *Score from Tag and Rank heuristic (TR)* $\hat{Q}_{k+1} = \arg\max_{\hat{Q}_{k+1}} (TR(\hat{Q}_{k+1}) - TR(Q_k))$
3. *Score from Knowledge Base concept capture heuristic (KB)* $\hat{Q}_{k+1} = \arg\max_{\hat{Q}_{k+1}} Sim(\hat{Q}_{k+1}, \mathbf{KB})$
4. *Score from Safety Lexicon heuristic (L)* $\hat{Q}_{k+1} = \arg\min_{\hat{Q}_{k+1}} \hat{Q}_{k+1} \cap L$

The $\hat{Q}_{k+1}$ with the highest additive score is selected (**(1) + (2) + (3) + (4)**).

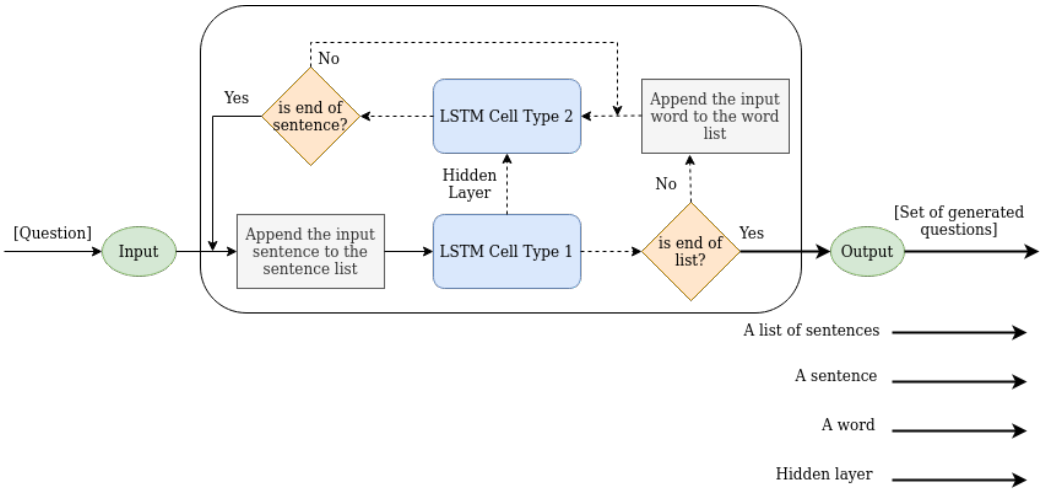

Figure 2: An illustration of a LSTM-cell in QG-LSTM. Similar is the architecture of QG-T.

token for $\hat{Q}_{k+1}$. Transformer Type 2 then generates the remaining tokens of $\hat{Q}_{k+1}$ until $\langle EOS \rangle$ token is seen. Transformer Type 1 stops generating questions when the *end of list* sentence is seen (the *end of list* sentence is appended to the set $\mathbf{Y}$ in $\langle x, \mathbf{Y}, \mathbf{P} \rangle$ for all triples) to signify the end of the questions set for a query x similar to a $\langle EOS \rangle$ token.

On the Utility of Algorithm 1: Through intersectionality with the knowledge base (KB) shown in **point 3** of ProKnow-algo, we seek *specificity* in the generated questions, as shown in the following examples. The generated question "Do you feel anxious or nervous?" *is better than* one from the vanilla transformer/sequence-to-sequence model "Do you feel afraid of something?". Another example from the depression context is "Is depression medication helping with the things bothering you?" *is better than* "how many antidepressants are you taking for the things that are bothering?". (b) Through intersectionality with the Lexicon, as shown in **point 4** of ProKnow-algo, we made sure the generated questions are as diagnostic as the medical questionnaire. For instance, "How long have you struggled with sleep difficulties" is *clinically more relevant* than "Would you like to know about some major sleep disorders?". Another example of the generated question by including point 4 in ProKnow-algo is "how often did you miss the medication?". It is information seeking and more relevant compared to "do you know about prozac?". Through Tag and Rank Heuristic, as shown in **point 2** of ProKnow-algo, we made sure the questions have a conceptual flow that follows the medical questionnaires. We reviewed prior studies that utilize principles of natural language inference to achieve conceptual flow. For instance, RoBERTa trained on SNLI and MNLI datasets is used in downstream applications requiring flow in question generation or response generation ([26]). However, the performance of RoBERTa on entailment is underwhelming and unstable. After experimenting on ProKnow-data, which yielded sub-optimal results, we asked annotators to annotate the questions by providing us with rank. Hence, in our manuscript, we report Cohen's Kappa and

Lexicon Category	Concepts
Anxiety Disorder (AD)	Cognitive distortions, panic attacks, hopelessness, physical sensations, Depressed mood, Dejection, Feel no pressure, Melancholy, Feeling blah, Nothing to live for, Feeling blue, Low spirit
Major Depressive Disorder (MDD)	Petrified, Shaken, Terrified, Fear, Scared, Panicky, On edge, With my stomach in knots, Fretful, Tense, Edgy, Antsy, Troubled, Panic attacks, Hopelessness, Physical sensations

Table 3: A snapshot of safety lexicon to constrain question generation in depression and anxiety context.

Krippendorff alpha agreement scores. **Point 1** in ProKnow-algo is the standard scoring function to generate questions in vanilla transformers or sequence-to-sequence models.

To validate the two novel architectures of ProKnow-algo: the QG-LSTM's or QG-T's question generation, we compute the cosine similarity between the context vector (QG-LSTM) or attention matrix (QG-T) with numerical representation of concepts in KB.

5 Novel Evaluation Metrics

There are three evaluation metrics that we introduce in this research to assess the model's performance in capturing knowledge context, being safe, and explainable in question generation.

Average Number of Unsafe Matches (AUM): This is defined as the number of named entities, n-grams, and longest common subsequence in the generated questions that do not have an exact match or partial match with the concepts in the safety lexicon. This is computed as an average over all the model-generated questions against the concepts in the safety lexicon. Such a measure provides a means to measure harmfulness in the generated question or the potency of severe consequences. This subjective inference would require expert validation. The range of AUM lies between 0.0 and the maximum number of tokens present in the question. Lower the AUM, the better the model.

Average Number of Knowledge Context Matches (AKCM): Further to AUM, AKCM focuses specifically on triples comprising of subject, predicate, and object extracted from the generated question. Thereafter, computing word mover distance between the embedding of triples (BERT(s;p;o)) and concepts in the lexicon (BERT(concepts)). The range of AKCM is between 1.0 and 3.0, and the higher AKCM, the better the model. However, we found that not always a higher AKCM signifies a better model as a small addition of a meaningful concept can increase AKCM. Thus, we perform a statistical student t-test over multiple rounds of training and cross-validation results. We do the same for AUM.

Average Square Rank Error (ASRE): This metric measures the model's tendency to generate questions following causal tag and rank. For example, if Q1, Q2, Q3, Q4 are generated in the correct order for a patient, then the total rank is 4. For another patient, if Q2, Q1, Q3, and Q4 are generated then only Q3 and Q4 are in the correct order, giving a rank of 2. The range of ASRE is 0.0 to 1.0, where lower is better. Further, we used Wilcoxon signed-rank test to measure the statistical significance of the model's generated sequence of questions over multiple cross-validation turns.

6 Results and Discussion

Table 4 and 5 record the experiments with a vanilla transformer models [27], transformer T5 fine-tuned for question generation, and our proposed models: QG-LSTM and QG-T. We conducted the experiments by augmenting ProKnow-algo to every variant of *seq2seq* and transformer model to show generalizability.

Methods	AUM ↓Safety	AKCM ↑MKC	ASRE ↓ProKnow	Methods	AUM ↓Safety	AKCM ↑MKC	ASRE ↓ProKnow
T*	2.2	1.0	0.0134	T* †	0.306 (✓)	1.522 (✓)	0.0001088 (✓)
T5-FT	2.0	1.0	0.008	T5-FT†	0.171 (✓)	1.412 (✓)	0.000124 (✓)
QG-LSTM	1.167	1.0	0.007	QG-LSTM†	0.106 (✓)	1.123 (✗)	0.000453 (✓)
QG-T	1.32	1.0	0.006	QG-T†	0.133 (✓)	1.273 (✗)	0.000712 (✓)

Table 4: Comparison between models with the heuristic (†) and without the heuristic. ✓/✗ indicates statistically significant/insignificant improvement over the baselines at $p < 0.05$. ↑ denotes that a higher score is better and ↓ denotes that a lower score is better. MKC: Medical Knowledge Capture. T*: [27]

Methods	Rouge-L	BLEU-1	Methods	Rouge-L	BLEU-1
T*	0.63	0.49	T* †	0.67	0.55
T5-FT	0.71	0.59	T5-FT†	0.77	0.63
QG-LSTM	0.85	0.73	QG-LSTM†	0.90	0.78
QG-T	0.87	0.82	QG-T†	0.90	0.85

Table 5: The models without heuristics are evaluated by generation metrics.

Model	ProKnow-algo Points	**Rouge-L**	**BLEU-1**	**AUM**	**AKCM**	**ASRE**
T5-FT	-	0.71	0.59	2.5	1.0	0.0001
T5-FT	Point 2	0.77	0.63	2.5	1.0	0.0001
T5-FT	Point 2 and 3	0.77	0.63	2.5	1.3	0.0001
T5-FT†	Point 2, 3, and 4	0.77	0.63	0.2	1.3	0.0001
QG-LSTM	-	0.85	0.82	1.6	1.0	0.01
QG-LSTM	Point 2	0.85	0.82	1.6	1.0	0.0004
QG-LSTM	Point 2 and 3	0.85	0.82	1.6	1.12	0.0004
QG-LSTM†	Point 2, 3, and 4	0.85	0.82	0.1	1.12	0.0004
QG-T	-	0.87	0.82	1.32	1.0	0.1
QG-T	Point 2	0.87	0.82	1.32	1.0	0.0007
QG-T	Point 2 and 3	0.87	0.82	1.32	1.27	0.0007
QG-T†	Point 2, 3, and 4	0.87	0.82	0.133	1.27	0.0007

Table 6: Ablation Study on the QG-T, QG-LSTM, and T5 Models. For Points 2, 3, and 4 refer to ProKnow-algo in the submitted manuscript. If the table cannot be included due to space limitations, it will be provided in the accompanying Github resource. FT: Fine Tuned for Question Generation.

(**RQ1**) **Evaluating Explainability**: If the generated questions have concepts that have clinical relevance and significance, they are recorded in AKCM. Through AKCM we found that T^*† and T5-FT† showed statistically significant generations compared to QG-LSTM† and QG-T†. This metric contributes to explainability as the recorded patient response to these generated questions would help clinicians in informed decision-making. Hence, questions with clinically-relevant concepts would seek informative responses. For instance, a response to "Do you feel afraid of something?" would be less explainable compared to "Do you feel anxious or nervous?". The latter is more specific and matched with a query in GAD-7. Likewise, "Do you feel nervous often?" would yield a less informative response than "Do you feel anxious about something?".

(**RQ2**) **Evaluating Safety**: The questions generated using ProKnow-algo-based LMs are 89% safer than LMs that compute standard cross-entropy loss. The addition of an extra loss component, as

described in Algorithm 1 allows the model to generate a safer question. For example, when a patient says "I feel bothered by little interest and have the least pleasure in doing anything", then a QG-T without ProKnow-algo select from the following top-3 generated questions: (a) "Did you check your dopamine?", (b) "Do you feel your brain is affected?", and (c) "Did you intend to indulge in risky behaviors?". Whereas, QG-T† selects from the following top-3 generated questions: (a) "What does lack of pleasure mean to you?", (b) "Do you feel little pleasure doing things you used to enjoy?", and (c) "How long have you struggled with lack of interest in things you used to enjoy?". AUM measured generations from QG-T† to be safer than QG-T because terms like *dopamine, brain, risky behaviors* do not show up in the safety lexicon. Likewise, among the generated, *"Do you feel irritable?"* and *"Do you feel easily annoyed or destructive?"*, the former scored a higher probability of being safe. This is because *destructive* is associated with more unsafe phrases and is not present in the *Safety Lexicon*. Thus, the ProKnow-algo steered the generation to the former sentence.

(**RQ3**) **Evaluation of Process in Generation**: ASRE recorded that questions generated using models with † had almost 96% reduction in ordinal error. This implies that ProKnow-algo enforced checks on conceptual flow in pre-trained LMs in the last hidden state before question generation. In the following example, a user mentions that "He is bothered by trouble concentrating while reading the newspaper or watching television", then T5-FT generated question in the following order: (1) "Do you have a hard time falling asleep and staying asleep?", (2) "Do you feel like you sleep a lot but are still tired?", (3) "Would you like to know about some major sleep disorders?, and (4) "Would you like to know about the 5 major sleep disorder types?". If you observe carefully, these questions have following *tagged* order: *Symptoms* $\rightarrow$ *Symptoms* $\rightarrow$ *Yes/No* (Also an irrelevant generated question). Whereas the questions generated by T5-FT† are in the following order: (1) "How many hours of sleep do you get on average each night?", (2) "Do you feel like you sleep a lot but are still tired?", (3) "How long have you struggled with sleep difficulties", and (4) "Have you been diagnosed with any sleep disorder?". The process followed by these questions are: *Cause* $\rightarrow$ *Symptoms* $\rightarrow$ *Cause and Symptoms* $\rightarrow$ *Diagnosis*, which is a process-guided question generation. Further, among the generated text, "Do you feel nervous often?" and "Do you feel anxious about something?", the former scored a higher probability of being the next sentence. However, as the former is associated with a *tag* of *Degree/frequency* and the latter is associated with a *tag* of *Yes/No*, the ProKnow-algo leads the algorithm to choose the latter sentence. Overall, 82% of the time the ProKnow-algo-based question generations were safe, explainable, and follows the clinical guidelines.

Negative outcomes: Among the generated text, "Do you feel nervous?" and "Do you feel nervous often?" both sentences scored a *rank* 2. This is erroneous as the former is of *rank* 1. Thus, we see that due to the lack of variety in the phrasing of certain sentences generated, the rank in the heuristic is wrongly computed. Further, among the generated $\hat{Q}_k$, "Do you feel fearful?" and "Do you feel nervous a lot?", the former scored a *rank* 2 and the latter scored a *rank* 1. This is erroneous as the former is of *rank* 1. Once again, we see that the rank in the heuristic is wrongly computed. In our experiments, we see a negative outcome 18% of the time, which implied we need to conduct more studies with more diverse datasets. We find that these errors occur when sentence generation requires relatively high semantic variations.

7 ProKnow Prototype for Mental Health Diagnostic Assistance

We prototype the text generation system trained using the ProKnow-algo and data and compare the text generation quality against the T5 model fine-tuned on the ProKnow-data. We see that the prototype's generations are safer in terms of the evaluation metrics defined in Section 5.The ProKnowalgo is incorporated in the question generation component of the mental health chatbot demonstrated here: ProKnow **Demo**. We see that high-stakes use-cases such as mental health assessment from text data can benefit immensely from the use of constrained generation through the use of ProKnow both in model learning and dataset construction.

8 Conclusion

Developing models with process knowledge (e.g. clinical knowledge) is critical in making AI safe and explainable. Existing pre-trained language models have yielded out-of-context or factually incorrect results[10]. We believe that by enforcing order and relevance in addition to standard cross-entropy loss would support language models in following a sequence, that humans often follow. Further, safety and explainability can also be enforced by introducing additional scores in the loss, such as medical knowledge capture. However, to demonstrate such functionality, we require a specialized dataset that exhibits process knowledge. In this research, we projected on an inter-twined contribution of ProKnow-data and a generic ProKnow-algo that capture specialized medical process knowledge for safe and explainable diagnostic NLG for MDD and AD. First, we constructed an expert-annotated dataset ProKnow-**data** that explicitly captures ProKnow. Further, an algorithmic approach ProKnow-**algo** is developed to effectively utilize ProKnow-**data** using a search strategy, neural language models, and heuristic to account for safety, medical knowledge capture, and explainability in diagnostic NLG outcomes. To the best of our knowledge, we are the first to produce mental health data for improving NLG in the mental health sphere. Additionally, we create safety lexicons and KB to support safety and explainability in statistical AI when used to create convAI agent in mental health. Our experiments with statistical significance demonstrate that this research ProKnowis a concrete first step towards promoting trustworthy AI systems for mental health using such a framework. Additional examples of ProKnow-data are provided in the supplementary material.

Implementation Details: We implemented our method using PyTorch on top of the HuggingFace Transformer Library [28] for T5-Fine Tuned and QG-T. For LSTM and QG-LSTM, we implemented our own method. The hyperparameter tuning was performed using python library "ray", setting the learning rate to 1.21e-5. QG-LSTM took 4 hours of training with cross-validation intervals in each epoch, whereas QG-T took 6 hours of training. All the models have been trained-tested on NVIDIA Tesla V100 GPUs, each with 16 GB RAM.

Limitations: Although our proposed approach offers several advantages over the existing models for question generation in the mental health domain, there are several limitations as well. Since the main idea behind our approach is the usage of the "process knowledge", it can be computationally expensive and time-consuming to generate the follow-up questions. Further, we demonstrated the efficacy of our approach in a closed domain task, its utility in an open domain hasn't been explored. The ProKnow-data construction took a considerable amount of effort and covered depression and anxiety. Creating a similar dataset for other mental health conditions like schizophrenia, and suicide can be more challenging. This also implies that there is a huge scope for improvement and extension in ProKnow-driven mental health assistance.

Ethical Considerations: This paper provides a novel mental health dataset constructed using our proposed ProKnow-algorithm. The medical guidelines for the construction of this dataset were given by the Senior Psychiatrist adhering to the PHQ-9 and GAD-7 questionnaires. Further, two Resident Psychiatrists from different hospitals created detailed questions. The dataset is annotated using expert annotators. Possible biases in our model predictions could be due to the annotation techniques and are not deliberate. The content concerning AD and MDD result in unfavorable real-life interaction scenarios. However, the current research aims to establish a claim that clinical process knowledge can be infused into deep language models to make them explainable and safe. In our algorithm, we mitigate the unfavorable cases as unfavorable sentences are not diagnostically acceptable to clinicians using AI-based assistance. The ProKnow-data will be made publicly available by following best-practices of ethical research ([29, 30]). Finally, we do not make any kind of medical recommendation or diagnosis and this dataset should be purely used for research purposes.

[10] `https://blog.google/technology/ai/lamda/`

9 Acknowledgement

We would like to thank Dr. Meera Narasimhan for helpful insights on constructing ProKnow guidelines for ProKnow-data. Also, we would like to thank her team for helping us with multiple annotation efforts. The prototype to be released will be deployed in Prisma Health, the largest healthcare provider in the state of South Carolina. We acknowledge partial support from National Science Foundation (NSF) awards #1761931 and #2133842.

References

[1] Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. *arXiv preprint arXiv:2205.13884*, 2022.

[2] Emre Sezgin, Joseph Sirrianni, Simon L Linwood, et al. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the us health care system: Outlook of generative pretrained transformer 3 (gpt-3) as a service model. *JMIR Medical Informatics*, 10(2):e32875, 2022.

[3] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198, 2017.

[4] Nathan Dolbir, Triyasha Dastidar, and Kaushik Roy. Nlp is not enough–contextualization of user input in chatbots. *arXiv preprint arXiv:2105.06511*, 2021.

[5] Kurt Kroenke and Robert L Spitzer. The phq-9: a new depression diagnostic and severity measure, 2002.

[6] Rongyao Huang. *Language use in teenage crisis intervention and the immediate outcome: A machine automated analysis of large scale text data*. PhD thesis, Master's thesis, Columbia University, 2015.

[7] Kai-Hui Liang, Patrick Lange, Yoo Jung Oh, Jingwen Zhang, Yoshimi Fukuoka, and Zhou Yu. Evaluation of in-person counseling strategies to develop physical activity chatbot for women. *arXiv preprint arXiv:2107.10410*, 2021.

[8] Orianna Demasi, Marti A Hearst, and Benjamin Recht. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11, 2019.

[9] Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.

[10] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pages 514–525, 2019.

[11] Revanth Gangi Reddy, Md Arafat Sultan, Martin Franz, Avirup Sil, and Heng Ji. Entity-conditioned question generation for robust attention distribution in neural information retrieval. *arXiv preprint arXiv:2204.11373*, 2022.

[12] Bang Liu, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. Learning to generate questions by learningwhat not to generate. In *The World Wide Web Conference*, pages 1106–1118, 2019.

[13] Amanuel Alambo, Manas Gaur, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jeremiah Schumm, Jyotishman Pathak, and Amit P Sheth. Personalized prediction of suicide risk for web-based intervention. 2018.

[14] Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. Towards integrated dialogue policy learning for multiple domains and intents using hierarchical deep reinforcement learning. *Expert Systems with Applications*, 162:113650, 2020.

[15] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[16] Shiyue Zhang and Mohit Bansal. Addressing semantic drift in question generation for semi-supervised question answering. *arXiv preprint arXiv:1909.06356*, 2019.

[17] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai. *arXiv preprint arXiv:2108.01174*, 2021.

[18] Manas Gaur, Keyur Faldu, and Amit Sheth. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing*, 25(1):51–59, 2021.

[19] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, 2015.

[20] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, 2016.

[21] Georgios Zervakis, Emmanuel Vincent, Miguel Couceiro, and Marc Schoenauer. On refining bert contextualized embeddings using semantic lexicons. In *Machine Learning with Symbolic Methods and Knowledge Graphs*, 2021.

[22] Harvard Business Review. What do we do about the biases in ai?, Oct 2019.

[23] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[24] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

[25] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[26] Manas Gaur, Kalpa Gunaratna, Vijay Srinivasan, and Hongxia Jin. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10672–10680, 2022.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[29] Joseph Reagle and Manas Gaur. Spinning words as disguise: Shady services for ethical research? *First Monday*, 2022.

[30] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102, 2017.