Sindbad at AraHealthQA Track 1: Leveraging Large Language Models for Mental Health QA

AbdulRahman A. Morsy[*1], **Saad Mankarious**[*1], **Aya Zirikly**[1,2]
[1]Department of Computer Science, The George Washington University (USA)
[2]Center for Language and Speech Processing (CLSP), Johns Hopkins University (USA)
{abdulrahman.morsy, saadm, ayah.zirikly}@gwu.edu

Abstract

Mental health detection in online discourse is a growing area of NLP research, particularly for low-resource languages such as Arabic, where stigma and limited access to professional care make anonymous, technology-driven solutions valuable. In the context of the AraHealth shared task, we were provided with three subtasks: multi-label classification for questions, multi-label classification for answers, and a QA system leveraging models developed in the previous two tasks. Our approach employed data augmentation to address class imbalance, as certain categories in the dataset were significantly underrepresented. Since our method relied on prompting models to classify questions and answers as well as to generate answers for the QA system, we utilized Gradient-free Edit-Based Instruction Search (GrIPS) to optimize prompt selection. Our system achieved strong results across all three subtasks, ranking 1st in answer classification and 3rd in both question classification and QA system answer generation.

1 Introduction

Mental health issues are a global concern with substantial economic and social impact (Santomauro et al., 2021). This challenge is particularly pronounced in Arabic-speaking communities, where discourse around mental health remains stigmatized, and access to professional resources is often limited or treated as a luxury (Khatib et al., 2023). Such constraints motivate NLP research that can detect and address mental health concerns using online data (Zirikly et al., 2019; Shing et al., 2018), enabling more anonymous, unrestricted, and accessible support tools for individuals in the Arab world (Hassib et al., 2022).

In this shared task (Alhuzali et al., 2025), we were provided with a dataset curated from an Arabic-language online forum that follows a question–answer (QA) pattern between patients and mental health professionals (Alhuzali et al., 2024). The dataset comprises 350 annotated instances, each containing a question, its corresponding answer, and categorical labels. Specifically, every sample is assigned one or more labels from seven possible question categories, as well as one or more labels from three possible answer categories, thereby constituting a multi-label classification setting. The distribution of these categories is highly imbalanced, especially for questions, which motivated our data augmentation approach. We employed GPT to generate additional samples for minority classes, resulting in a more balanced dataset. Using the augmented data, we performed instruction fine-tuning with a range of pre-trained models. In parallel, we explored few-shot prompting for multi-label classification in Task 1 (question classification) and Task 2 (answer classification). For Task 3 (QA system), we leveraged the fine-tuned models from the first two tasks and applied Gradient-free Edit-Based Instruction Search (GrIPS) to optimize the prompts used for answer generation.

Our system achieved strong results across all tasks: we ranked 1st in answer classification (Sub-Task 2) and 3rd in both question classification (Sub-Task 1) and QA answer generation (Sub-Task 3).

2 Background

The shared task (Alhuzali et al., 2025) focused on three subtasks in the Mental Health track: (1) multi-label question classification, (2) multi-label answer classification, and (3) a QA system for generating appropriate answers using models from the first two tasks. Each instance in the dataset consists of a question, its corresponding answer, a question category (one of seven possible labels), and an answer category (one of three possible labels). For

[*]Equal contribution

Proceedings of The Third Arabic Natural Language Processing Conference, pages 184–191
November 8-9, 2025 ©2025 Association for Computational Linguistics

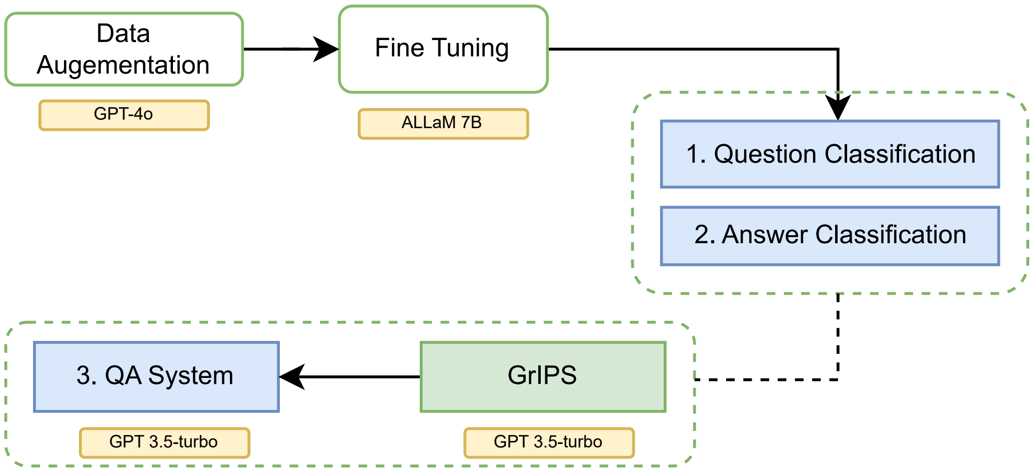

Figure 1: Overall pipeline for our approach. Beginning with the raw data, we generate synthetic samples and leverage them to perform classification and then answer generation.

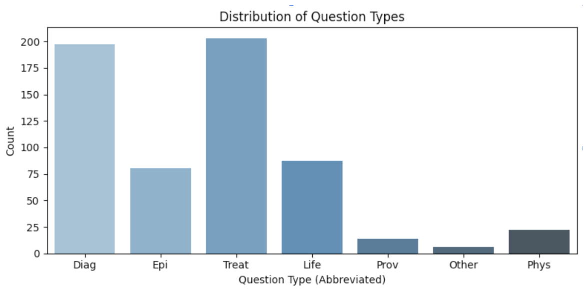

Figure 2: Question class distribution, showing significant imbalance between categories.

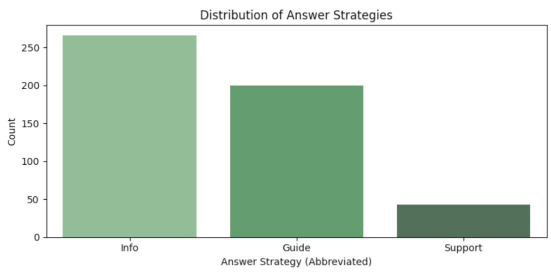

Figure 3: Answer class distribution.

example, a question about treatment options for depression could be labeled under *Treatment* for questions and *Supportive Advice* for answers.

The question taxonomy spans seven categories, covering clinical reasoning and practical guidance: Diagnosis (A) for interpreting findings, Treatment (B) for therapeutic options, Anatomy and Physiology (C) for biomedical knowledge, Epidemiology (D) for disease progression and causes, Healthy Lifestyle (E) for wellness habits, Provider Choices (F) for healthcare navigation, and Other (Z) for miscellaneous queries. Answer strategies fall into three broad types: Information (1) delivering factual content and resources, Direct Guidance (2) offering actionable recommendations, and Emotional Support (3) providing reassurance or encouragement (Alhuzali et al., 2024).

The dataset contains 350 labeled QA pairs from an Arabic-language mental health forum. Figures 2 and 3 show the category distributions. The question categories are heavily imbalanced, with certain categories having fewer than 10 samples, while the largest category has over 175 samples. The answer categories are also imbalanced, though less severely. This imbalance strongly motivated our data augmentation approach to generate synthetic samples for minority classes.

Our participation covered all three tracks, and our contribution is novel in its integration of prompt optimization (via GrIPS) with both few-shot prompting and instruction fine-tuning for imbalanced and low-resource Arabic mental health classification tasks. Related work in Arabic NLP has explored mental health (Alhuzali and Alasmari, 2025, 2024), but to our knowledge, no prior shared task system has combined prompt optimization with synthetic minority oversampling for both classification and answer generation in this domain. For instance, MedArabiQ (Abu Daoud et al., 2025) introduced a benchmark for evaluating large language models on Arabic medical tasks, covering a wide range of QA problems.

3 System Overview

We built the QA system by leveraging our models from Subtasks 1 and 2, which classify questions and answers. As shown in Figure 4, we appended the predicted question category to the QA prompt. We used the question category predicted labels (obtained from fine-tuning) as guidelines for the QA model. Furthermore, the system prompt is optimized using GrIPS as Section 3.3 explains.

3.1 Data Augmentation

To address dataset size constraints and class imbalance, we employed GPT-4o (temperature = 0.7) to synthesize additional training instances. Augmentation targeted the least frequent labels: (3) and its multi-label variants (1,3), (2,3), and (1,2,3) for Subtask 2; and labels C, D, E, and F for Subtask 1.

Prompt construction incorporated: (i) role specification to enforce domain-appropriate tone; (ii) category definitions from (Alhuzali et al., 2024); (iii) in-context exemplars; (iv) explicit formatting constraints (e.g., fixed sample counts, variable lengths); and (v) lexical variation controls to minimize redundancy.

For Subtask 1, we generated 300 synthetic samples (50 each for D and E; 100 each for C and F), expanding the dataset from 350 to 650 samples. For Subtask 2, we generated 160 samples (40 per target configuration), expanding the dataset from 350 to 510 instances.

Following the generation of each set of samples, we performed a human evaluation to assess the quality and relevance of the generated samples. For this purpose, we randomly selected approximately one-third of each set for detailed inspection with respect to fluency, label relevance, and adherence to the specified constraints.

Appendix A.1 includes examples of both the prompts and the generated data. Additionally, the full datasets, including all generated samples, are available on our GitHub repository.[1]

3.2 Model Fine-Tuning

We used fine-tuned models from Subtask 1 (Question Classification) to explicitly solve Task 1. Furthermore, we leveraged the best performing model (ALLaM 7B) to develop the QA system as shown in Figure 1. Given a question that the system needs to respond to, we obtain a predicted category label from the fine-tuned models that we then provide in the prompt. This is the second step in our prompt building process in Figure 4.

3.3 Prompt Optimization with GrIPS

Gradient-free Instructional Prompt Search (GrIPS) is a technique proposed by (Prasad et al., 2022) to efficiently optimize the prompts used for our QA system. We used GrIPS to optimize the system prompt of our QA System as Figure 4 shows. Our implementation of GrIPS follows an iterative prompt optimization process. Starting from an initial prompt, we generate candidate variations through targeted mutations, such as structural adjustments, content refinements, and cultural adaptations for six iterations. Each candidate prompt is evaluated on a subset of the training data using BERTScore F1 to measure the alignment between the model-generated and reference answers. See Figure 5 for BERTScore performance for each iteration. The highest-scoring prompt is retained for the next iteration, and the process is repeated for a fixed number of optimization rounds. This approach enables systematic improvement of prompt effectiveness without gradient-based updates, ultimately yielding an optimized instruction that enhances model performance on the QA system.

4 Experimental Setup

4.1 Sub-Task 1: Question Classification

Prior to building the QA system, we employed fine-tuning to classify questions. We experimented with an array of PLMs and used the best performing model (ALLaM 7B) in the QA System. The hyperparameters used are listed in Table 1. Table **??** in the appendix also shows the performance on the evaluation portion (20 percent) of our augmented dataset.

4.2 Sub-Task 2: Answer Classification

For Subtask 2, we fine-tuned **MARBERT** (Abdul-Mageed et al., 2021) on the augmented dataset described in Section 3.1 for multi-label classification, using the hyperparameters in Table 2, consistent with (Alhuzali and Alasmari, 2025) for the same task. Model training employed a maximum sequence length of 256, the *Binary Cross-Entropy* loss function, and an *AdamW* optimizer.

An 80/20 train-validation split was used for cross-validation to ensure that model performance was stable and not driven by outliers. Then, we

[1] https://github.com/AbdulRahmanBenatia/Sindbad-AraMentalQA-SharedTask

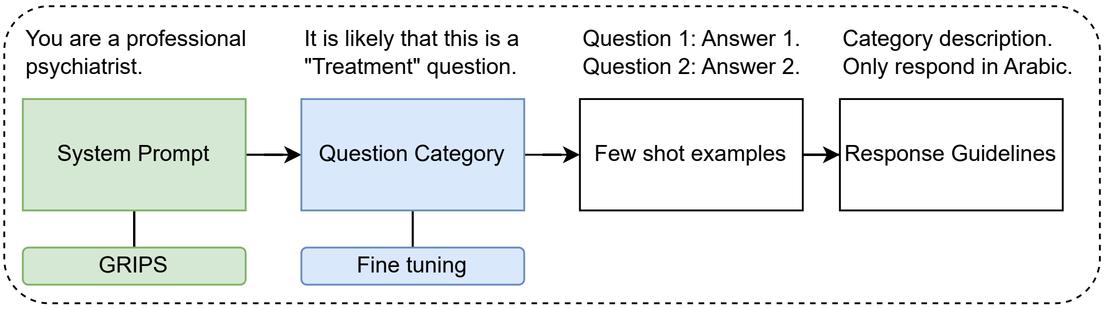

Figure 4: Overall prompt structure for the QA System.

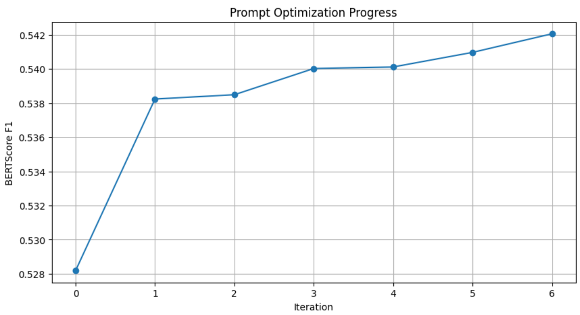

Figure 5: GPT-3.5-turbo Performance during GrIPS prompt optimization.

Parameter	Value
Learning rate	1×10^{-4}
Batch size	2
Gradient accumulation steps	4
Epochs	3
Weight decay	0.01
Early-stop patience	2
LORA rank (r)	8
LORA alpha	16
LORA dropout	0.05
Quantization	None
Optimizer	adamw_torch

Table 1: Default hyperparameters for fine-tuning on Subtask 1, question classification.

re-trained the model on the full augmented set and used it to generate predictions for the official test set. We report our results using the following evaluation metrics: *Weighted-F1* and *Jaccard* score, as recommended by the shared task organizers.

4.3 Sub-Task 3: Question Answering System

We employed `gpt-3.5-turbo` via the OpenAI API for question–answer (QA) generation. The system adopted a few-shot prompting approach with three manually carefully selected examples that represent diversity in target labels. Categories were drawn from a predefined taxonomy (A–F, Z) covering diagnosis, treatment, anatomy/physiology, epidemiology, healthy lifestyle, provider choice, and miscellaneous queries.

Prompts were structured into: (i) a system role enforcing professional and empathetic Arabic medical responses; (ii) category-specific contextual descriptions; (iii) explicit response guidelines; and (iv) optional in-context examples. The temperature parameter was fixed at 0.0 to ensure deterministic output and reproducibility.

Parameter	Value
Hidden size	768
Batch size	8
Dropout	0.1
Early-stop patience	10
Epochs	15
Learning rate	2×10^{-5}
Optimizer	AdamW

Table 2: Hyperparameters for MARBERT fine-tuning.

5 Results

5.1 Sub-Task 1

Our ALLaM-based system achieved third place in Subtask 1 with *Weighted-F1* $= 0.53$ and *Jaccard* $= 0.49$ as shown in Table 3.

Weighted-F1	Jaccard	Ranking
0.53	0.49	3rd

Table 3: Results on the official test set for Subtask 1.

5.2 Sub-Task 2

As shown in Table 4, our system achieved *Weighted-F1* $= 0.79$ and *Jaccard* $= 0.71$, ranking first in Subtask 2. The observed performance gains were primarily attributed to the targeted augmentation of under-represented label combinations.

Weighted-F1	Jaccard	Ranking
0.79	0.71	1st

Table 4: Results on the official test set for Subtask 2.

5.3 Sub-Task 3

On the official test set, our `gpt-3.5-turbo` system achieved a BERTScore of 0.668, ranking 3rd in Subtask 3. This performance reflects the benefit of structured, category-aware prompting and few-shot exemplars, though the gap to the top systems suggests potential for further domain adaptation.

BERTScore	Ranking
0.668	3rd

Table 5: Results on the official test set for Subtask 3.

6 Conclusion

In this work, we developed a system for Arabic mental health QA tasks that integrates instruction fine-tuning on augmented data, few-shot prompting, and gradient-free prompt optimization via GrIPS. Our approach effectively addresses class imbalance and low-resource challenges, achieving first place in answer classification (Sub-Task 2) and third place in both question classification (Sub-Task 1) and QA answer generation (Sub-Task 3). These results demonstrate the effectiveness of combining synthetic data generation, fine-tuning, and prompt engineering to enhance large language model performance in specialized, low-resource domains and open the way for further research addressing Arabic mental health NLP.

References

M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Virtual.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks. *arXiv e-prints*, pages arXiv–2505.

H. Alhuzali and A. Alasmari. 2025. Pre-trained language models for mental health: An empirical study on arabic q&a classification. *Healthcare*, 13(9):985.

H. Alhuzali, A. Alasmari, and H. Alsaleh. 2024. MentalQA: An annotated arabic corpus for questions and answers of mental healthcare. *IEEE Access*.

Hassan Alhuzali and Ashwag Alasmari. 2024. Evaluating the effectiveness of the foundational models for q&a classification in mental health care. *arXiv preprint arXiv:2406.15966*.

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash, and Leen Kharouf. 2025. Arahealthqa 2025 shared task description paper. In *Proceedings of ArabicNLP 2025*.

Mariam Hassib, Nancy Hossam, Jolie Sameh, and Marwan Torki. 2022. AraDepSu: Detecting depression and suicidal ideation in Arabic tweets using transformers. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 302–311, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hania El Khatib, Aisha Alyafei, and Madiha Shaikh. 2023. Understanding experiences of mental health help-seeking in arab populations around the world: A systematic review and narrative synthesis. *BMC Psychiatry*, 23(1):324.

Archiki Prasad, Pratyush Venkatesh, Yixin Xu, Rui Zhang, Junyi Jessy Li, Surya Kallumadi, and Bill Yuchen Lin. 2022. GrIPS: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Damian F. Santomauro, Ana M. Mantilla Herrera, Junfang Shadid, Peng Zheng, Charlie Ashbaugh, Giorgia Pigott, Harish G. Sheena, and et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 Data Augmentation

Figure 6 presents an example prompt along with three randomly selected generated responses, illustrating Answer augmentations associated with labels (1, 3).

Similarly, Figure 7 presents an example prompt together with five randomly selected generated responses, illustrating Question augmentations associated with label E.

A.2 Fine-tuning

We experimented with the following arrays of models during fine tuning:

```
"meta-llama/Llama-2-13b-chat-hf",
"ALLaM-AI/ALLaM-7B-Instruct-preview",
"silma-ai/SILMA-9B-Instruct-v1.0",
"aubmindlab/bert-base-arabertv2",
"UBC-NLP/MARBERT",
"CAMeL-Lab/bert-base-arabic-camelbert-
    mix"
```

Model	F1 Micro	F1 Wtd.	Jaccard
meta-llama	0.62	0.57	0.47
ALLaM-7B	**0.68**	**0.59**	**0.54**
SILMA-9B	0.62	0.57	0.48
arabertv2	0.68	0.59	0.53
MARBERT	0.68	0.59	0.53
camelbert-mix	0.68	0.59	0.53

Table 6: Performance of different models on the evaluation set for Subtask 1: Question Classification.

A.3 GrIPS

To optimize the prompt used in the QA system, we ran the optimization script against the initial template, which resulted in the final prompt header

```
#Initial Prompt Header
"You are an expert Arabic mental health
    assistant.
    Provide accurate, helpful responses
        to mental health questions in
        Arabic.
    Be professional yet empathetic in
        your answers."

#Final Prompt Header
"You are an expert Arabic psychiatric
    conditions assistant.
    Provide accurate, evidence-based
        responses to psychiatric
        conditions questions in Arabic.
    Be professional yet empathetic in
        your answers. Answer concisely
        in Arabic medical terminology.
        Answer concisely in Arabic
        medical terminology."
```

Note: this represents the prompt header, often supplied as a system propmpt to applicable models. In the rest of the prompt, we append the categories description as given the description of the task, along with other information to control the model response.

أنت طبيب نفسي بارع ذو شهرة واسعة وتحب إجابة أسئلة المرضى بدقة وخبرة.

لديك ثلاثة أنواع من الإجابات على أسئلة المرضى، منها ما يلي:

Information. This category includes answers that provide information, resources, etc. It also includes requests for information.

Emotional Support. This category includes answers that provide approval, reassurance, or other forms of emotional support.

أمثلة: واضح انك توتري قوي حاولى انك وأنت بتتكلمى مع الآخرين انك لا تأخذي الموضوع على انك في وضع تقييم ثقي في نفسك و ركزي عليها اكتر من رأي الناس فيك

سلامتك الوسواس مرض نفسي قابل للعلاج وكلما كان مبكرا اصبح ايسر في العلاج

انت حاسس ان ده افضل حل حاليا لان مفيش حاجة حلوة في حياتك وكل يوم بتعيشه مؤلم بالنسبة لك، طيب لو في فرصة تطلب المساعدة وتلاقي حاجات حلوة واهداف مهمة في حياتك، مش يمكن ده يخليك تعيد النظر في اختيارك، ارجو التواصل معايا عن طريق رقم التلفون يمكن اقدر اساعدك

في حالات اضطراب ثنائي القطب لا ينصح باستخدام مضادات الاكتئاب بشكل عام (مودابكس،سيربتاس)..الافضل الالتزام بخطتك العلاجية من قبل طبيبك المعالج.و في حال ظهور اي اعراض للانتكاسة مراجعة الطبيب للتقييم و اجراء اللازم…الف سلامة عليك.

هل يمكنك إعطاء الكثير من النماذج لهذا تقتصر على الدعم والمعلومات فقط دون إرشاد علاجي. فقط دعم ومعلومات بخبرة الطبيب النفسي البارع صاحب اللغة الواضحة.

رجاء أعطني على الأقل 40 إجابة من هذا النوع لأسئلة مختلفة، وليقتصر ردك على الإجابات مرقمة من 1 إلى 40 وبينها سطر فارغ.

اجعل طول الإجابات متنوعًا أي لا تقتصر على الجمل القصيرة أو الطويلة بل أريد تنوعًا واضحا.

(a) Example prompt.

- واضح من كلامك إنك تعاني من توتر مزمن. التوتر له تأثير كبير على الجسم والنفس، ويمكن أن يظهر في شكل صداع، أرق، أو حتى آلام عضلية. مجرد إدراكك للمشكلة هو خطوة كبيرة نحو التحسن. لا تحمل نفسك فوق طاقتها، وأنت تستحق أن ترتاح.
- الوسواس القهري مرض نفسي شائع، وله أسباب بيولوجية ونفسية. يتميز بأفكار متكررة وسلوكيات قهرية يصعب التحكم بها. وجودك هنا وسؤالك يدل على وعي ورغبة في التغيير، وهذا شيء مشجع جدًا.
- كثير من الناس يمرون بما تمر به، والشعور بالحزن لفترات طويلة قد يكون مرتبطًا بأعراض اكتئابية. ما تحسه حقيقي، وليس مبالغة، وأنت تستحق أن يُفهم ويُحتوى.

(b) Sample responses.

Figure 6: Augmentation prompt and sample generated responses for Answers (label 1, 3).

أنت مولع بعلم النفس والطب النفسي، تقرأ كثيرًا أسئلة المرضى
وإجابات الأطباء عليها. أنواع الأسئلة متعددة، منها النوع الآتي:

Healthy lifestyle: Questions are specified to diet, exercise, mood control and other lifestyle factors that can affect health.

أمثلة:

انا غير راضية عن نفسي من نايحة المظهر اصبحت لدي عقدة اشعر اني
بشعة والله انا اخجل من نفسي اصبحت دائما اقارن نفسي بالبنات الاخريات
واحسدهن على جمالهن كثيرا ما تأحسر على مضهري انعدمت ثقتي بنفسي
ودائما انضر الى المراة مع العلم اني لست بدالك السوء انا عادية
اغلب الناس من يقول اني جميلة لكن لا اصدقهم مع العلم انه ليس لدي
اي تشوهات معدا مستوى الاسنان احينا ما اتعرض لسخرية من اخوتي والله
اشعر بالاحباط شديد

انا رجل ابلغ من العمر اربعون عام ولم اتزوج من قبل ولكن الان افكر
في الزواج والبعض ينصحني بعدم الزواج لتقدم سني واني لا استطيع ان
اعيش حياة سعيده بعد هذا العمر

حزن الرغبه فنوم لفتره طويله لا اريد التكلم والحديث

هل يمكنك إعطائي 50 سؤالًا آخر من هذا النوع. رجاء التزم هذا النوع،
مع التنوع في الموضوعات وطول السؤال فبعض الأسئلة قصير وبعضها طويل
وهكذا.

ليس ضروريًّا أن تكون جميعها أسئلة، فبعضها أسئلة، وبعضها شكوى
يشتكيها المريض (مع التزام النوع).

(a) Example prompt.

- فقدت رغبتي في الخروج من المنزل وأصبحت أفضل العزلة طوال الوقت.
- أشعر أنني أتناول الطعام طوال الوقت دون سبب حقيقي، فقط للهروب من التوتر.
- هل المشي اليومي كافٍ لتحسين المزاج والنوم؟
- لم أعد أستمتع بشيء، حتى الهوايات التي كنت أحبها أصبحت مملة.
- هل قلة شرب الماء تؤثر على التركيز والنشاط؟

(b) Sample responses.

Figure 7: Augmentation prompt and sample generated responses for Questions (label E).