MILD Bot: Multidisciplinary Childhood Cancer Survivor Question-Answering Bot

Mirae Kim[1], Kyubum Hwang[1], Hayoung Oh[1†]
Min Ah Kim[2], Chaerim Park[2], Yehwi Park[2], Chungyeon Lee[2]
[1]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, South Korea
[2]Department of Social Welfare, Sungkyunkwan University, Seoul, South Korea
{miraekiim, bany1111}@g.skku.edu rimzzang123@gmail.com
{hyoh79, minahkim, skalxk, dalcong7930}@skku.edu

Abstract

This study introduces a Multidisciplinary chILDhood cancer survivor question-answering (MILD) bot designed to support childhood cancer survivors facing diverse challenges in their survivorship journey. In South Korea, a shortage of experts equipped to address these unique concerns comprehensively leaves survivors with limited access to reliable information. To bridge this gap, our MILD bot employs a dual-component model featuring an intent classifier and a semantic textual similarity model. The intent classifier first analyzes the user's query to identify the underlying intent and match it with the most suitable expert who can provide advice. Then, the semantic textual similarity model identifies questions in a predefined dataset that closely align with the user's query, ensuring the delivery of relevant responses. This proposed framework shows significant promise in offering timely, accurate, and high-quality information, effectively addressing a critical need for support among childhood cancer survivors.

1 Introduction

In recent decades, there have seen remarkable advancements in pediatric cancer survival rates, encompassing cancers diagnosed in children and adolescents aged 0 to 19 years (Siegel et al., 2024). Today, nearly 80% of these children achieve long-term survivor (Argenziano et al., 2023). Similarly, South Korea has achieved an impressive average 5-year survival rate (2017-2021) for childhood cancer, reaching 86.5% (Korea Central Cancer Registry., 2023). However, the growing number of survivors highlights the need to address their complex psychological and social needs (Choi, 2018; Lim, 2020). South Korea still lacks a comprehensive system for providing necessary psychosocial support (Kim et al., 2021), in contrast to the more developed systems in the United States (Kim et al., 2018). Furthermore, survivors often face challenges in accessing support services due to fears of disclosing their medical history and associated stigma, complicating their adjustment and well-being (Kim & Yi, 2012; Yi et al., 2014; Lown et al., 2015; Prasad & Goswami, 2021).

The COVID-19 pandemic has accelerated the adoption of digital health technologies, including conversational agents, in oncological care (Briggs et al., 2022). These technologies are now crucial for cancer screening, patient education, symptom monitoring, and psychological support. Notable examples include ChemoFreebot, which aids breast cancer patients in self-care (Tawfik et al., 2023), and Vivibot, which helps young adult cancer survivors manage anxiety (Greer et al., 2019). Despite their benefits, Wang et al. (2023) found that the use of conversational agents in cancer care remains limited, especially for childhood cancer survivors.

To address the critical gap in support for childhood cancer survivors, we propose a multidisciplinary question-answering (QA) bot

† Corresponding Author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 665–676

named the Multidisciplinary chILDhood cancer survivor question-answering bot (MILD). This bot offers a stigma-free platform for accessing information, resources, and emotional support. Using data from 860 academic articles, 2 publications, 18,565 news articles, 23 credible social media platforms, and 25 YouTube videos, we utilized OpenAI's GPT-4 Turbo model to generate precise responses. Our MILD bot framework includes an intent classifier to understand user queries and a semantic textual similarity (STS) model to fetch relevant answers from a pre-established database. The main contributions of this paper are:

- **Development of a MILD bot**: Introducing the first QA bot tailored to offer diverse expert responses for childhood cancer survivors.
- **Construction of a domain-specific STS dataset**: Enhancing our response model with a specialized STS dataset.
- **Evaluation with childhood cancer survivors**: Testing the MILD bot with real users demonstrated its effectiveness and potential for real-life application.

2 Background

The core mechanism of the MILD bot for childhood cancer survivors is an STS model, which retrieves relevant answers. To enhance its performance, we investigated several Korean STS datasets for fine-tuning, as summarized in Table 1.

Korean STS: Developed by Kakao Brain[‡], this dataset features 8,628 sentence pairs created using round-trip translation from the English STS dataset. Sentences are labeled for similarity on a scale from 0 (no similarity) to 5 (high similarity) (Ham et al., 2020).

KLUE STS: Part of the Korean Language Understanding Evaluation (KLUE) benchmark, this dataset includes 12,187 sentence pairs from practical contexts such as Airbnb reviews and news articles, providing a broad representation of real-world language use (Park et al., 2021).

Question pair v2: Curated from a non-domain specific online site, this dataset contains 6,888 sentence pairs with binary labels: 0 for dissimilar sentences and 1 for similar sentences.

Name	Language	Sentence pairs	Label
Korean STS	English Korean	8,628	0-5
KLUE STS	English Korean	12,187	0-5
Question pair v2	Korean	6,888	0,1
ParaKQC	Korean	10,000	None

Table 1: Korean STS datasets.

ParaKQC: The Parallel Korean Questions and Commands (ParaKQC) dataset includes 100 sets of 10 sentence pairs each, focusing on sentences with the same topic and intention. It does not use similarity labels, emphasizing parallel topics and intentions instead (Cho et al., 2020).

Despite the availability of multiple Korean STS datasets, their limited size poses a challenge to significantly improving model performance. Moreover, none of these datasets pertains to the domain of childhood cancer, limiting their effectiveness in enhancing the model's accuracy in this area. Consequently, in Section 3, we explore the development of a domain-specific STS dataset tailored to our model's needs.

3 Datasets for Childhood Cancer Survivors

In our study, we utilized various datasets to develop the MILD bot model, including both the benchmark dataset discussed in Section 2 and new datasets created specifically for our research. Table 2 provides an overview of all the datasets used. The new datasets are divided into two parts: training and inference QA. The training datasets are used for domain-adaptive training and fine-tuning of the models, while the inference QA dataset serves as a predefined database for the model to find the most appropriate answers.

3.1 Training Dataset

Childhood Cancer Survivor Domain Corpus: The Childhood Cancer Survivor (CCS) Domain Corpus was collected to enhance domain-adaptive training for our retrieval-based response model, which often struggles in untrained domains. Domain-adaptive training has been shown to significantly improve model performance

[‡] https://www.kakaobrain.com/

	Original	New			
Purpose	Training	Training		Inference QA	
Name	KLUE (STS / NLI)	Childhood Cancer Survivor Domain Corpus	Childhood Cancer Survivor STS	Expert QA	Peer Survivor QA
Source	Airbnb, News	860 Academic articles / papers, 2 Publications, 18,565 News articles	Inference QA	23 Social media platforms, Online survey, GPT-4 Turbo	40 Academic articles / papers, 25 YouTube videos
Size	40,185 (sp*)	2.2GB	31,456 (sp)	3,500 (qa**)	1,238 (qa)
				3,500 (qa)	

sp*: Sentence Pairs qa**: Question-and-answer Pairs

Table 2: Datasets for childhood cancer survivors.

(Gururangan et al., 2020), and even a small corpus can aid model specialization (Sanchez & Zhang, 2022). Additionally, Yao et al. (2021) noted that expanding the vocabulary with domain-specific terms can boost performance when resources are limited.

Childhood cancer survivors face concerns spanning medicine, law, and finance (Erdmann et al., 2021; Hendriks et al., 2021). To improve model quality, we collected 860 academic articles, papers, and 2 publications using keywords recommended by a psychosocial expert, such as "childhood cancer," "childhood leukemia," "childhood brain tumor," and "pediatric oncology." We also included online materials like news articles and posts related to childhood cancer, resulting in a corpus totaling 2.2GB. Despite our efforts to gather diverse sources, the specificity of the domain made it challenging to amass a large corpus. Therefore, following Yao et al. (2021)'s approach, we expanded the vocabulary with frequently occurring terms from the CCS Domain Corpus. More details can be found in Appendix A.

Childhood Cancer Survivor STS: The CCS Semantic Textual Similarity (STS) dataset was developed to fine-tune pretrained models for improved performance in STS tasks within the childhood cancer domain. As we mentioned in Section 2, existing Korean STS datasets are small and lack relevance to childhood cancer, posing challenges for model training (Ban, 2021). To address this gap, we created a new dataset tailored to this domain, leveraging an inference QA dataset with 3,500 questions covering childhood cancer survivor concerns, inspired by Thakur et al. (2021).

Figure 1 illustrates the overall process. We followed a three-step process:

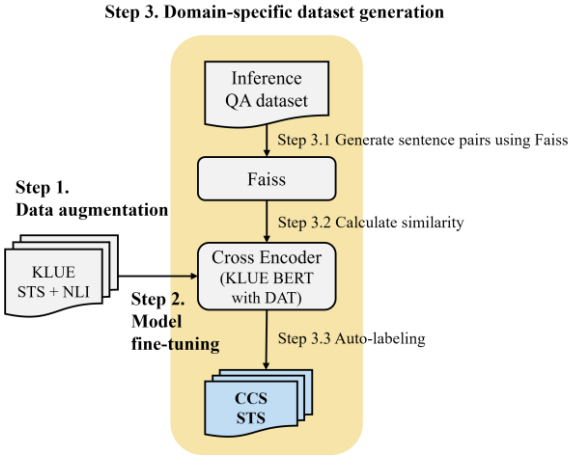

Figure 1: Process for creating the childhood cancer survivor STS dataset.

- **Step 1. Data augmentation:** To address the lack of a Korean STS dataset, we applied Gao et al. (2021)'s method, which demonstrated that Natural Language Inference (NLI) datasets can effectively enhance STS datasets. In our experiments, we utilized a Korean NLI dataset in three ways: (1) using "entailment" pairs as positive and "contradiction" pairs as negative, achieving a score of 0.8679; (2) assigning random similarity scores (0-1) to "contradiction" pairs, resulting in 0.8499; and (3) using only "entailment" pairs as positive, yielding the highest score of 0.8683. Notably, this third approach outperformed

	Overall Quality	Factuality	Completeness	Ease of Understanding	Empathy
Experts	**4.98 (0.72)**	**4.99 (0.73)**	**4.84 (0.75)**	4.90 (0.77)	**4.84 (0.87)**
Childhood Cancer Survivors	4.36 (0.63)	4.57 (1.22)	4.64 (1.15)	**5.43 (0.65)**	4.5 (1.09)

Table 3: Evaluation results of chatbot responses by experts and childhood cancer survivors (Numbers in parentheses indicate the standard deviation).

the baseline STS dataset score of 0.8657. Based on these findings, we transformed the NLI dataset from the KLUE benchmark into an STS dataset by treating "entailment" labeled instances as positive pairs. We then merged the KLUE STS and NLI dataset to further improve performance.

- **Step 2. Model fine-tuning:** We fine-tuned a cross-encoder model with the merged dataset. This model, pretrained on the CCS Domain Corpus, ensured a better understanding of the specific domain.

- **Step 3. Domain-specific dataset generation:** Using the Faiss model (Johnson et al., 2017), we generated similar sentence pairs from the inference QA dataset to create a domain-specific dataset. For each question, Faiss identified nine similar questions, excluding the original question. Each pair was auto-labeled using a fine-tuned cross-encoder model. The resulting dataset consists of 31,456 sentence pairs, all related to the childhood cancer domain, providing a substantial amount of data to train the model.

3.2 Inference QA Dataset

The inference QA dataset serves as the predefined database that our MILD bot model uses to select the most suitable responses. This dataset includes two question-and-answer pair databases (see examples in Appendix Table 7).

Expert QA: The Expert QA dataset includes opinions and solutions from three main expert groups: pediatric oncologists, social workers, and psychological and mental health professionals, focusing on childhood cancer survivors. We collected inquiries from 23 social media platforms, including the Korea Childhood Leukemia Foundation[§], Korea Association for Children with Leukemia and Cancer[**], and the National Cancer Information Center [††]. Additionally, an online survey with 119 childhood cancer survivors provided 1,283 genuine questions, reflecting their true concerns. All survey questions received IRB approval from the University Ethics Committee.

We used GPT-4 Turbo to generate responses, assigning it the roles of the selected experts. Eleven experts (6 social workers, 3 psychological and mental health professionals, and 2 pediatric oncologists) evaluated the responses on overall quality, factuality, completeness, ease of understanding, and empathy (Xu et al., 2023) using a 6-point Likert scale (Chomeya, 2010). Table 3 presents the evaluation scores. On average, the responses scored above 4 points across all aspects, indicating that GPT-4 Turbo's answers are highly valuable for childhood cancer survivors. To further assess the quality of GPT-4 Turbo's responses, we compared them with answers collected from 23 social media platforms. Experts consistently preferred GPT-4 Turbo's responses due to their informational richness and generally longer, more comprehensive style.

In cases where the GPT-4 Turbo responses were found unsatisfactory, experts provided gold-standard responses to ensure accuracy and completeness. We then refined the responses by incorporating expert feedback and applying few-shot prompting techniques (Brown et al., 2020), using real expert responses as references.

Peer Survivor QA: The Peer Survivor QA dataset is a key contribution, providing answers from actual peer survivors to childhood cancer survivors. The online survey revealed that participants preferred and felt greater empathy

[§] https://www.kclf.org/en/
[**] https://www.soaam.or.kr/english/
[††] https://www.cancer.go.kr/

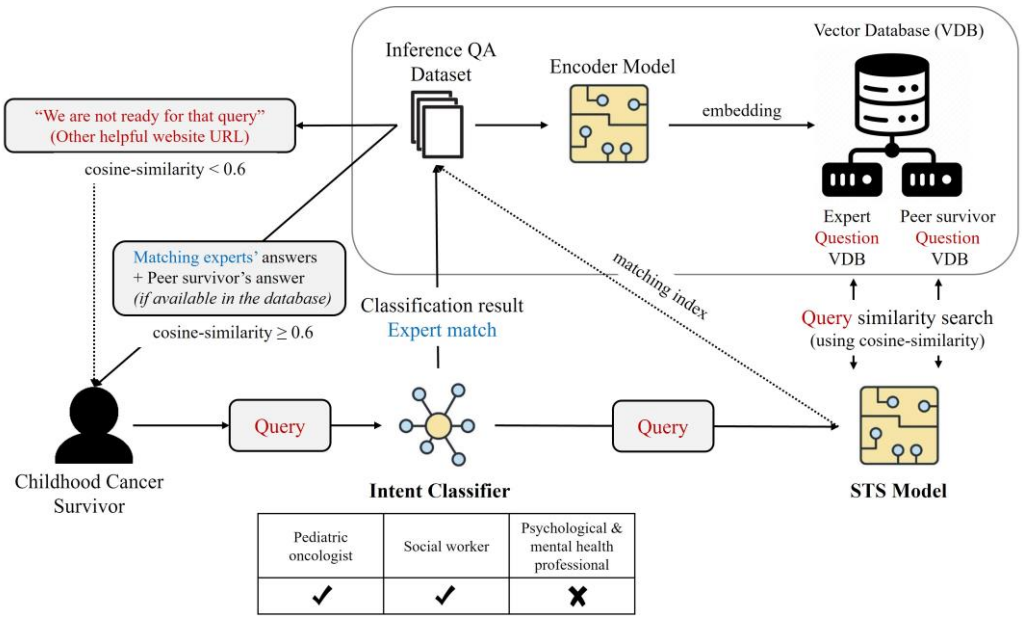

Figure 2: MILD Bot framework. The MILD Bot operates only if a similar question exists in the database. If the user's query is not relevant to the predefined database, the MILD Bot refrains from answering and instead responds with, "We are not ready for that query," to maintain accuracy.

from peer survivors' responses. Although these answers may lack detailed information, they offer valuable insights and experiences.

To gather authentic utterances from peer survivors, we meticulously extracted real interview responses from 40 academic articles and 25 YouTube videos. Our efforts focused on aggregating and anonymizing these utterances to ensure both authenticity and privacy. This dataset includes 1,283 responses from the 3,500 questions in the inference QA dataset, as we could not obtain peer survivors' answers for all questions.

4 The Proposed Scheme

The MILD bot consists of two main components. First, an intent classifier identifies the intention behind the survivors' queries. Then, an STS model matches the query with the most relevant question from the Expert QA and Peer Survivor QA datasets. Based on survivors' preferences, the system retrieves the most appropriate response. Figure 2 illustrates the overall architecture of our MILD bot model.

4.1 Intent Classifier

To address survivors' questions, we developed an intent classifier to match each query with the appropriate expert. The first step involved creating a training dataset. A psychosocial expert categorized sample questions in the inference QA dataset into three groups: pediatric oncologists, social workers, or psychological and mental health professionals. Questions relevant to multiple groups were assigned to all applicable groups for comprehensive responses.

This categorized dataset was used as a few-shot learning sample for GPT-4 Turbo, creating a cost-effective training set. According to the recent studies, GPT-4 excels in natural language reasoning tasks (Liu et al., 2023; Gilardi et al., 2023). After generating a labeled dataset with GPT-4 Turbo, we fine-tuned our intent classifier. The results are detailed in Section 5.

Although the inference QA dataset for the MILD bot includes responses from experts across all domains for every question, providing all answers simultaneously can overwhelm childhood cancer survivors. Additionally, not all questions require responses from every expert; for instance, detailed medical inquiries do not need input from social workers or mental health professionals. On the other hand, questions that involve multiple areas of expertise—such as those requiring empathetic support—benefit from responses from all relevant sources, including peer survivors. With the help of an intent classifier, the MILD bot ensures that users receive comprehensive, informative, and empathetic responses tailored to their specific needs.

Setting	Cosine-pearson	Euclidean-pearson	Dot-pearson
KLUE BERT	0.77	0.81	0.75
KLUE BERT w/ DAT*	**0.86**	**0.88**	**0.87**
KLUE BERT w/ Expanded Vocab** / DAT	0.86	0.88	0.86
KLUE BERT w/ Expanded Vocab / DAT / CCS STS***	**0.93**	**0.91**	**0.90**

w/ DAT*: with Domain Adaptive Training
w/ Expanded Vocab**: Pretraining on Expanded Vocabulary
w / CCS STS***: Fine-tuning with CCS STS Dataset

Table 5: Evaluation metrics of domain-adaptive training.

4.2 Domain-specific STS Model

STS quantitatively measures the semantic similarity between texts (Yang et al., 2020). We used the STS task to match survivors' questions with those in the inference QA dataset to find the most semantically similar questions. This helps identify the most appropriate answers within a constrained source environment.

While a bi-encoder architecture is generally less precise than a cross-encoder, it offers faster response times and requires fewer resources (Reimers & Gurevych, 2019). To enhance performance, we applied domain-adaptive training and fine-tuned the model using a targeted STS dataset, as detailed in Section 3. Upon receiving a query, the model retrieves one or two answers from the inference QA dataset based on its intent. If the dataset lacks questions similar to the user's query, the MILD bot avoids providing an answer rather than offering the closet match. After evaluating various thresholds, we found that a similarity score of 0.6 optimizes the bot's performance.

5 Experiment

We conducted experiments to enhance our MILD bot's performance, evaluating different Korean pretrained language models for optimal training efficiency.

First, we chose KcBERT (Lee, 2020) for its robust handling of typological errors, given its pretraining on a large online corpus. Second, we selected KM-BERT (Kim et al., 2022), pretrained on a Korean medical corpus, to better understand medical terminology. Finally, we chose KLUE BERT (Kim et al., 2023) for its superior performance among Korean BERT models.

We fine-tuned the intent classifier using these models to assess their impact on performance. Our evaluation included domain-adaptive training and an ablation study to refine its effectiveness. Additionally, we conducted a human evaluation with 14 childhood cancer survivors who interacted with our MILD bot.

5.1 Training an Intent Classifier

As described in Section 4, we developed a cost-effective multi-label dataset using GPT-4 Turbo to automatically label the inference QA dataset. We tested various Korean BERT models to identify the best performer. The dataset, comprising 3,500 questions, was split into training, validation, and testing sets in a 7:1.5:1.5 ratio.

To evaluate multi-label performance, we calculated the Exact-Match Ratio (EMR) and label-based weighted scores for precision, recall, and F1-score. The results, shown in Table 4, indicate that the KLUE BERT model outperformed the others.

Korean BERT	EMR	Precision	Recall	F1-score
KcBERT	0.72	0.87	0.89	0.88
KM-BERT	0.74	0.89	0.89	0.89
KLUE BERT	**0.76**	**0.90**	**0.90**	**0.90**

Table 4: Evaluation metrics of an intent classifier.

5.2 Domain-adaptive Training

Based on the findings from Section 5.1, we selected KLUE BERT for domain-adaptive training. We conducted an ablation study with four scenarios to demonstrate its effectiveness:

- Using the pretrained KLUE BERT as a baseline.
- Pretraining KLUE BERT with a CCS Domain Corpus.
- Pretraining KLUE BERT with an expanded vocabulary and the CCS Domain Corpus.
- Pretraining KLUE BERT with both an expanded vocabulary and the domain-specific corpus, further fine-tuned using the CCS STS dataset.

To assess performance in the STS task, we compiled a test dataset combining sentence pairs from KLUE STS, KLUE NLI, and CCS STS datasets. We trained the model with a learning rate of 0.05 using the AdamW optimizer over 5 epochs. The results, shown in Table 5, indicate that domain-adaptive training is particularly effective for the childhood cancer domain. Adding 294 words to KLUE BERT's existing 32,000-word vocabulary did not significantly impact performance. Notably, the creation of a domain-specific STS dataset significantly improved performance.

5.3 Human Evaluation

We evaluated the MILD Bot with 14 participants aged 20 to 41, all of whom had been diagnosed with cancer during their childhood or adolescence in South Korea, using the ngrok service (see Appendix C for details). To mitigate potential risks, survivors under the age of 20 were excluded from the study. Over two weeks, each participant engaged with the MILD bot in at least 10 sessions, each lasting over 15 minutes and involving 10 to 20 questions per session. In the final session, participants completed an online survey using the same evaluation criteria as the expert evaluation. The results are shown in Table 3. Participants rated their overall satisfaction and the usefulness of the MILD bot an average of 3.78 out of 5 points. Moreover, all participants expressed a desire to reuse the MILD bot.

6 Conclusion

We developed the MILD bot, a multidisciplinary question-answering bot specifically for childhood cancer survivors. Based on survey findings, we prioritized providing accurate information with empathetic tones. To build the bot, we gathered diverse data from academic articles, social media, YouTube, news sources, and peer survivors' utterances. Using these datasets, we performed domain adaptive training on the KLUE BERT model to enhance MILD bot's understanding of relevant information. The MILD bot features an intent classifier to identify query intentions and an STS model to retrieve and provide the most appropriate answers from a predefined database, ensuring precise information tailored to the needs of childhood cancer survivors.

7 Limitations and Future Works

While our study demonstrates the MILD bot's effectiveness in the childhood cancer domain, we identified some limitations in both the dataset and the model. The lack of a comprehensive CCS Domain Corpus limits the model's performance. Despite our efforts, the domain-specific data remains insufficient, with our corpus size at 2.2GB, relatively small compared to other studies (Chalkidis et al., 2020; Rasmy et al., 2021; Syed & Chung, 2021). Future research will focus on augmenting the dataset with translated English-language data. Moreover, testing with 14 childhood cancer survivors revealed the need for medical expertise in addressing specific questions. We plan to expand the dataset with more cancer-related information, guided by pediatric oncologists.

Although Retrieval-Augmented Generation (RAG) systems are widely adopted in modern QA tasks, we chose a retrieval-only approach. Since this is the first attempt to develop an informational chatbot specifically for childhood cancer survivors, we adopted a conservative stance to ensure user safety, given the sensitivity of the target population. While RAG systems offer advantages, they can also generate incorrect answers when retrieved resources contain conflicting information (Barnett et al., 2024, Feldman et al., 2024). After expanding the dataset to cover a broader range of questions, we plan to compare the performance of a RAG-based system with our current retrieval-only approach.

Ethics Statement

This research was approved by University Ethic Committee (IRB. No. 2023-11-049).

Acknowledgments

This Paper was supported by AI Convergence Research Fund, Sungkyunkwan University, 2023, the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1F1A1074696), BK21 FOUR Project (Bigdata Research and Education Group for Enhancing Social Connectedness Thorough Advanced Data Technology and Interaction Science Research,5199990913845) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2023-00254129, Graduate School of Metaverse Convergence (Sungkyunkwan University)), Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of game-based digital Therapeutics technology for adolescent mental health (psychological and behavioral control) management, Project Number: RS-2024-00344893 and the MSIT (Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00459638) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. 2024. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1), 12–49. https://doi.org/10.3322/caac.21820

Maura Argenziano, Alessandra Di Paola, and Francesca Rossi. 2023. Childhood Cancer Survivors: An Overview of the Management of Late Effects. *Cancers*, 15(12), 3150. https://doi.org/10.3390/cancers15123150

Korea Central Cancer Registry, National Cancer Center. Annual report of cancer statistics in Korea in 2021, Ministry of Health and Welfare, 2023

Kwonho Choi. 2018. Vulnerabilities and Psychosocial Service Needs of Childhood Cancer Survivors and their Caregivers Based on the Cancer Trajectory. *Health and Social Welfare Review*, 38(2), 417–451. https://doi.org/10.15709/HSWR.2018.38.2.417

Su-Jin Lim. 2020. Review of Childhood Cancer Survivors' Health-related Need. *Journal of the Korea Convergence Society*, 11(3), 361–368. https://doi.org/10.15207/JKCS.2020.11.3.361

Min-Ah Kim, Kwonho Choi, and Jung-won You. 2021. Current Psychosocial Care for Children and Adolescents with Cancer and Their Families: Perspectives of Service Providers in Hospitals and Community Welfare Agencies. *Health and Social Welfare Review*, 41(3), 130-159. https://doi.org/10.15709/hswr.2021.41.3.130

Min-Ah Kim, Jaehee Yi, and Kwonho Choi. 2018. Perceived benefits and challenges of psychosocial service uses for adolescents and young survivors of childhood cancer. *Health and Social Welfare Review*, 38(3):247–278. https://doi.org/10.15709/HSWR.2018.38.3.247

Min-Ah Kim and Jaehee Yi. 2012. Childhood Cancer Survivor's Services Needs for the Better Quality of Life. Journal of Korean Academy of Child Health Nursing, *Korean Academy of Child Health Nursing*. https://doi.org/10.4094/jkachn.2012.18.1.19

Jahee Yi, Min-Ah Kim, and Jungsu Kim. 2014. Stigma Experiences and Psychosocial Responses to Stigma among Childhood Cancer Survivors. *Mental Health and Social Work*.

E. Anne Lown, Farya Philliphs, Lisa A. Schwartz, Abby R. Rosenberg, and Barbara Johnes. 2015. Psychosocial Follow-Up in Survivorship as a Standard of Care in Pediatric Oncology. *Pediatric blood & cancer*, 62 Suppl 5(Suppl 5), S514–S584. https://doi.org/10.1002/pbc.25783

Maya Prasad and Savita Goswami. 2021. Barriers to long-term follow-up in adolescent and young adult survivors of childhood cancer: Perspectives from a low-middle income setting. *Pediatric blood & cancer*, 68(12), e29248. https://doi.org/10.1002/pbc.29248

Logan G Briggs, Muhieddine Labban, Khalid Alkhatib, David-Dan Nguyen, Alexander P Cole, and Quoc-Dien Trinh. 2022. Digital technologies in cancer care: a review from the clinician's perspective. *Journal of comparative effectiveness research*, 11(7), 533–544. https://doi.org/10.2217/cer-2021-0263

Elham Tawfik, Eman Ghallab, and Amel Moustafa. 2023. A nurse versus a chatbot – the effect of an empowerment program on chemotherapy-related side effects and the self-care behaviors of women living with breast Cancer: a randomized controlled trial. *BMC nursing*, 22(1), 102. https://doi.org/10.1186/s12912-023-01243-7

Stephanie Greer, Danielle Ramo, Yin-Juei Chang, Michael Fu, Judith Moskowitz, and Jana Haritatos. 2019. Use of the Chatbot "Vivibot" to Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer Treatment: Randomized Controlled Feasibility Trial. *JMIR mHealth and uHealth*, 7(10), e15018. https://doi.org/10.2196/15018

Alexandar Wang, Zhiyu Qian, Logan Briggs, Alexander P Cole, Leonardo O Reis, and Quoc-Dien Trinh. 2023. The Use of Chatbots in Oncological Care: A Narrative Review. *International journal of general medicine*, 16, 1591–1602. https://doi.org/10.2147/IJGM.S408208

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 422–430, Online. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean Language Understanding Evaluation. *ArXiv*, https://doi.org/10.48550/arXiv.2105.09680

Won Ik Cho, Jong In Kim, Young Ki Moon, and Nam Soo Kim. 2020. Discourse Component to Sentence (DC2S): An Efficient Human-Aided Construction of Paraphrase and Sentence Similarity Dataset. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6819–6826, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Chris Sanchez and Zheyu Zhang. 2022. The Effects of In-domain Corpus Size on pre-training BERT. *ArXiv*, https://doi.org/10.48550/arXiv.2212.07914

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-Distill: Developing Small, Fast and Effective Pretrained Language Models for Domains. *ArXiv*, https://doi.org/10.48550/arXiv.2106.13474

Friederike Erdmann, Line Elmerdahl Frederiksen, Audrey Bonaventure, Luzius Mader, Henrik Hasle, Leslie L Robison, and Jeanette Falck Winther. 2021. Childhood cancer: Survival, treatment modalities, late effects and improvements over time. *Cancer epidemiology*, 71(Pt B), 101733. https://doi.org/10.1016/j.canep.2020.101733

Manya Jerina Hendriks, Erika Harju, Katharina Roser, Marcello Ienca, and Gisela Michel. 2021. The long shadow of childhood cancer: a qualitative study on insurance hardship among survivors of childhood cancer. *BMC Health Services Research*, 21, 503. https://doi.org/10.1186/s12913-021-06543-9

Byunghyun Ban. 2021. A Survey on Awesome Korean NLP Datasets. *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 1615-1620. https://doi.org/10.1109/ICTC55196.2022.9952930

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. *ArXiv*, https://doi.org/10.48550/arXiv.2010.08240

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, Hervé Jégou. 2017. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7, 535-547. https://doi.org/10.48550/arXiv.1702.08734

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. *ArXiv*, https://doi.org/10.48550/arXiv.2305.18201

Rungson Chomeya. 2010. Quality of Psychology Test Between Likert Scale 5 and 6 Points. Journal of Social Sciences, 6(3), 399-403. https://doi.org/10.3844/jssp.2010.399.403

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

Shot Learners. *ArXiv*, https://doi.org/10.48550/arXiv.2005.14165

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *ArXiv*, https://doi.org/10.48550/arXiv.2304.03439

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120. https://doi.org/10.1073/pnas.2305016120

Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, and Yonghui Wu. 2020. Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models. *JMIR medical informatics*, 8(11), e19735. https://doi.org/10.2196/197354

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Conference on Empirical Methods in Natural Language Processing*.

Junbum Lee. 2020. Kcbert: Korean comments bert. *In Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*. 437-440.

Yoojoong Kim, Jong-Ho Kim, Jeong Moon Lee, Moon Joung Jang, Yun Jin Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joon Joo, and Sanghoun Song. 2022. A pre-trained BERT for Korean medical natural language processing. *Scientific reports*, 12(1), 13847. https://doi.org/10.1038/s41598-022-17806-8

Mirae Kim, Kyubum Hwang, Hayoung Oh, Heejin Kim, and Min Ah Kim. (2023). Can a Chatbot be Useful in Childhood Cancer Survivorship? Development of a Chatbot for Survivors of Childhood Cancer. *In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. 4018–4022. https://doi.org/10.1145/3583780.3615234

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4, 86. https://doi.org/10.1038/s41746-021-00455-y

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *In Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Muzamil Hussain Syed and Sun-Tae Chung. 2021. MenuNER: Domain-Adapted BERT Based NER Approach for a Domain with Limited Dataset and Its Application to Food Menu Domain. *Applied Sciences*, 11(13), 6007; https://doi.org/10.3390/app11136007

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*. 194-199. https://doi.org/10.48550/arXiv.2401.05856

Philip Feldman, James R. Foulds, Shimel Pan. 2024. RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots. *ArXiv, abs/2403.01193*. https://doi.org/10.48550/arXiv.2403.01193

Inference QA Dataset				
Question	Expert QA			Peer Survivor QA
	Pediatric Oncologist	Social worker	Psychological and Mental Health Professionals	Peer Survivor
How should I take care of my health after recovering from cancer?	Managing health is very important. Even after recovery, it is necessary to regularly check for recurrences…	Childhood cancer survivors may need special health care due to the effects of their treatment….	After recovering from the cancer, it is most important to regularly check both physical and mental health. …	A survivor visited a Cancer Survivor Support Center and consulted a counselor on how to manage their health. …
Should I talk about my cancer experience with others?	Deciding whether to share your cancer experience with others is entirely a personal choice. …	… Sharing your story can have different significance for each person. Some may choose to share their experience …	… However, it's crucial to remember that the decision is entirely yours. Your willingness to share…	Individuals who had experienced childhood cancer mentioned that they felt the need to disclose their experience …

Table 7: Examples of inference QA dataset

A Expanded Vocabulary

To select the domain-specific vocabulary and avoid out-of-vocabulary issues, we followed these steps: We aimed to extract only relevant data from the CCS domain corpus by counting the frequency of each noun and including new nouns only if they occurred more than 100 times. This process automatically excluded authors' names and paper-related words. Given the original vocabulary size of KLUE BERT, which includes 32,000 words, we ultimately added only 294 new words. Figure 3 illustrates the overall process of adding new vocabulary, and Table 6 provides samples of the expanded vocabulary.

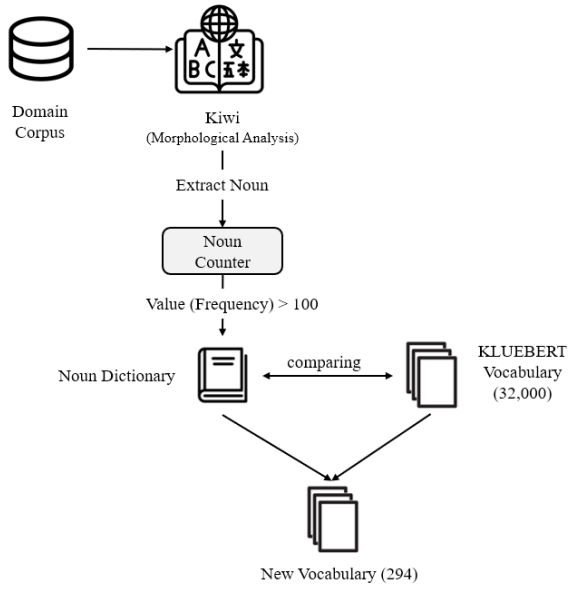

Figure 3: Vocabulary Expansion Process

Expanded Vocabulary		
Sick children	Erythrosis	Brain tumor
Antigen	Leukocyte	Platelet
Serum	Wilms Cancer	School age
Stomatitis	Alopecia	Childhood Cancer

Table 6: Samples of expanded vocabulary

B Examples of Datasets

Among the 3,500 pairs in the inference QA dataset, the Expert QA subset included 3,500 question-and-answer pairs, while the Peer Survivor QA subset comprises 1,283 question-and-answer pairs. Questions in both datasets were collected from diverse sources as mentioned in Section 3, but the responses were sourced differently. In the Expert QA, all responses were generated by GPT-4 Turbo. Although we initially collected responses from various sources, we generated new responses based on these originals to incorporate expert's tone and empathetic nuances. Furthermore, for questions from the online survey, we could not collect responses.

In contrast, the Peer Survivor QA dataset features responses collected directly from sources, reflecting real experiences of peer survivors. To avoid generating potentially inaccurate or fabricated responses, we deliberately chose not to generate these responses. Table 7 provides samples of question-and-answer pairs from each dataset. Given that GPT-4 Turbo responses tend to be detailed and lengthy, we abbreviated them in Table 7 to highlight difference between each expert.

C Testing of the MILD Bot

For testing our MILD bot, we created a temporary web-based bot service using ngrok, allowing 14 participants to easily access the MILD bot via a provided URL. During the test, we collected each question they asked and immediately aggregated them into the original dataset. Table 8 shows samples of their questions.

	Questions
1	How many days does it take for the blood type to change after all allogeneic hematopoietic stem cell transplant?
2	Why is the strongest chemotherapy administered for 7 days before a bone marrow transplant?
3	Why does my spine hurt so much after receiving an immune-boosting injection?

Table 8: Sample questions

Figure 4 shows the main web GUI participants encountered when they accessed the URL. The main web page displayed the following message:
"*Hello. I am the bot here to help with questions from childhood survivors. During our conversation, you can ask anything related to childhood cancer. Each response will include input from various expert groups (pediatric oncologist, social worker, psychological and mental health professional) or peer survivors. You have 15 minutes to freely ask your questions. When you want to end the conversation, type 'end' in the chat box. Let's begin. Before starting, please enter your 4-digit identification number (numbers only, no spaces).*"

Figure 4: MILD bot main web GUI