Leveraging Pre-Trained Language Models to Streamline Natural Language Interaction for Self-Tracking

Young-Ho Kim¶   **Sungdong Kim**¶   *__Minsuk Chang__¶   **Sang-Woo Lee**¶‡

¶NAVER AI Lab, ‡NAVER CLOVA

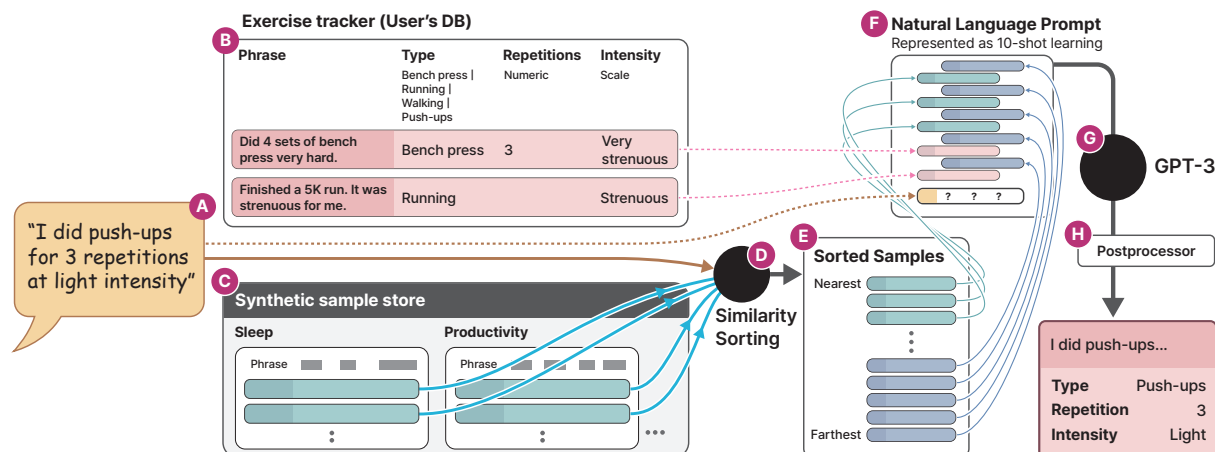{ygho.kim, sungdong.kim, minsuk.chang, sang.woo.lee}@navercorp.com

Figure 1: A two-shot learning example on our NLU framework, when the person has two prior items for exercise tracker **B** . The system augments the prompt with synthetic seed samples **C** to transform the task into a 10-shot learning problem **F** . Refer to Appendix A for an example prompt.

Abstract

Current natural language interaction for self-tracking tools largely depends on bespoke implementation optimized for a specific tracking theme and data format, which is neither generalizable nor scalable to a tremendous design space of self-tracking. However, training machine learning models in the context of self-tracking is challenging due to the wide variety of tracking topics and data formats. In this paper, we propose a novel NLP task for self-tracking that extracts close- and open-ended information from a retrospective activity log described as a plain text, and a domain-agnostic, GPT-3-based NLU framework that performs this task. The framework augments the prompt using synthetic samples to transform the task into 10-shot learning, to address a cold-start problem in bootstrapping a new tracking topic. Our preliminary evaluation suggests that our approach significantly outperforms the baseline QA models. Going further, we discuss future application domains toward which the NLP and HCI researchers can collaborate.

1 Introduction

Self-tracking tools (*e.g.*, mHealth apps like Fitbit [Fitbit, Inc. 2021a]) help people longitudinally track their health and activity in a structured and systematic manner. The advancement of Natural Language Interaction (NLI) techniques have opened new opportunities for designing novel self-tracking systems with which people can intuitively record their data using speech and/or chat; specifying long and complex information in natural language is generally more flexible and expressive than using predetermined forms in traditional graphical widgets (Kim et al., 2021; Luo et al., 2021, 2020). As a result, there is a growing interest in building *speech-mediated self-tracking* tools to offer low-burden (*e.g.*, Luo et al. 2021) and accessible (*e.g.*, Kim et al. 2022) self-tracking.

Yet, existing systems predominantly incorporate bespoke (and mostly rule-based) natural language understanding (NLU) logics optimized for capturing uniform information in a specific tracking theme, compromising generalizability and scalability to diverse user inputs and contexts. For example, the NLU of Data@Hand, a visual exploration app for fitness data, is implemented using syntax-based rules with POS (part of speech) tags and predefined

---

keywords (Kim et al., 2021). Such an approach is known to be vulnerable to a selection of vocabulary and individualized linguistic patterns (Kim et al., 2019a, 2021). Furthermore, extending the NLU to support a new type of data (e.g., exercise sessions) is a demanding task because it requires appending new rules manually.

Despite the promise of deep-learning-based NLP approaches, it is still challenging to develop flexible and scalable NLU models for self-tracking mainly because the design space of self-tracking is broad in terms of topics and data formats (Epstein et al., 2020; Kim et al., 2017). Hence, it is overwhelming to collect natural language datasets that cover the entire space and varied domains.

In this work, we introduce a novel NLP task for self-tracking focusing on in-situ data collection scenarios where people capture their retrospective activity logs. We also propose a novel NLU framework (Figure 1) that supports this task, which incorporates GPT-3 (Brown et al., 2020), a large-scale pre-trained language model (PLM), to handle linguistic variations of natural language commands (*e.g.*, "*I drank a cup of coffee an hour ago*" or "*At 3:00 PM, had an Americano.*"). To bootstrap in-context learning on GPT-3, the framework leverages synthetic samples constructed from simulations with 24 seed tracking schemas. Given a natural language phrase, the framework extracts values for data fields from a data table. The phrase may be terse and specify only a subset of data fields.

Our preliminary evaluation shows that our prompt augmentation approach using synthetic seed samples was effective in extracting appropriate information from the input phrases in a low-resource scenario. In pure zero-shot cases, GPT-3 underperformed the T5-based model (Lin et al., 2021; Raffel et al., 2020). However, it outperformed when augmented with synthetic seed samples that had different data schemas and tracking topics. The performance increased by the number of prior examples in the corresponding data schema but saturated with more examples. Our findings demonstrate the opportunities of GPT-3's in-context learning abilities for avoiding a cold start problem of the natural-language-based data collection task.

2 Background

Self-tracking is a powerful means of understanding oneself and self-promoting positive behavior

changes (Choe et al., 2014; Li et al., 2010). People capture their activities in a variety of topics including but not limited to physical/mental health, finance, productivity, diet, and sleep (Epstein et al., 2020). The data points collected for self-tracking usually describe a phenomenon during a time interval or for an associated time point at a specific granularity like minute or day (Kim et al., 2019b). The phenomenon information consists of various types of data fields such as numbers (*e.g.*, step count, heart rate), texts (*e.g.*, description of a stress episode), scales (*e.g.*, productivity, stress level, sleep quality) or choices (*e.g.*, type of mood) (Jeon, 2016; Kim et al., 2017).

While fitness trackers can capture various health metrics, many of the human activities cannot be captured by sensors and still require a manual input to be captured. As an effective way to reduce the manual input burden, the speech modality has recently gained interest and was applied to smart speakers (*e.g.*, Fitbit Skill [Fitbit, Inc. 2021b], MyFitnessPal Skill [Under Armour, Inc 2021]) and research prototypes (*e.g.*, ModEat [M. Silva and A. Epstein 2021], TandemTrack [Luo et al. 2020], FoodScrap [Luo et al. 2021]). These tools support speech-based data capture through smart speakers, smartwatches, or smartphones. Our work expands this growing body of speech-based self-tracking research by proposing a unified framework of NLU to support flexible phrasing of multifaceted information in arbitrary tracking topics, which are not yet supported by prior systems.

3 NLU Framework for Self-Tracking

3.1 Task Description

Imagine a person uses a self-tracking platform that consists of multiple data tables for sub-topics. This is analogous to common health platforms such as Fitbit (Fitbit, Inc., 2021a) and Apple Health (Apple Inc., 2021) with multiple data tables for step count, body weights, food, or water intake. We refer to the individual data tables as **trackers** (*e.g.*, Ⓑ in Figure 1), and data points for each tracker as **items** (*e.g.*, colored bars in Ⓑ in Figure 1). A tracker comprises multiple input fields with six data types—*number*, *Likert scale*, *single-choice*, *multiple-choice*, *short-form text*, and *long-form text*, which are derived from prominent data types of existing self-tracking apps (Jeon, 2016; Kim et al., 2017). The person may use natural language to insert a new item to the database. For exam-

ple, he or she may speak, "*I did* **push-ups** *for* **three** *repetitions at* **light** *intensity,*" to describe an item, *{Exercise → push-ups, Repetitions → 3, Intensity → light }*, upon finishing the exercise session. Such an interaction may be performed via a smartphone app, chatbots, or voice assistants. This main task of our framework can be represented as $itm'_{trk} = [v^0_{trk}...v^i_{trk}...v^{n-1}_{trk}] = NLU(trk, phr)$, where NLU derives a list of value v for n fields of the tracker trk from the phrase phr. Assuming that we have little or no instances for Itm'_{trk}, we solve $NLU()$ through few-shot learning with GPT-3. We turned to PLM because it can be switched to a different problem upon a new natural language prompt with a handful of examples (Liu et al., 2021b), whereas traditional machine learning models require a large amount of task-specific datasets.

3.2 Prompt Augmentation

In the early stage of using trk, the person may not have little or no items for it; the system does not have sample instances with the same data schema to be put in a prompt for few-shot learning, when it receives a new phrase for trk. To overcome the instability of accuracy in low-shot cases (Brown et al., 2020), we transformed the NLP task into a 10-shot learning problem by augmenting the model prompt (*c.f.*, Appendix A) using *synthetic samples*. We constructed a synthetic sample store (C in Figure 1) with 504 item-phrase pairs from 24 trackers (21 pairs per each tracker). The trackers were manually composed by the authors (see Appendix B for an exhaustive list of trackers). We randomly generated item samples and phrases that describe the content using GPT-3. Four authors iteratively inspected the data and corrected wrong matches between the values and the phrase. Each sample contains a *subset* of data fields of the tracker, to simulate the cases when people do not include all field values in a single utterance. (See the second item in B in Figure 1 that omitted *Repetitions*.)

The current implementation mixes both the nearest five and the farthest five samples in a prompt (D – F in Figure 1), inspired by Liu et al. 2021a and Zhao et al. 2021. We used cosine similarity between the embeddings calculated using a sentence transformer `multi-qa-MiniLM-L6-cos-v1` in the `sentence-transformers`[1] package. When there exist items and phrases for the tracker, they are treated as the nearest samples and placed

near the output of the prompt (F in Figure 1). The framework passes the prompt to GPT-3 via OpenAI's API[2]. Specifically, we used `text-davinci-002`, the most capable *InstructGPT* (Ouyang et al., 2022) model optimized for following human prompts. Finally, the postprocessor (H in Figure 1) parses the plain text output into a data table and matches the choice labels to the nearest ones in a tracker schema using the same transformer used in D.

4 Preliminary Evaluation

To obtain preliminary insights on the feasibility of our approach, we evaluated the task outcomes from a series of scenarios, using the synthetic samples as a validation dataset.

Baselines We evaluated *TransferQA* and pure zero-shot *GPT-3* as two baseline models. Since the proposed task has not been thoroughly explored in the NLP discipline, we chose TransferQA, one of the best-performing model for *dialog state tracking*, whose task is the most similar to the proposed one. TransferQA is a T5-based model pre-trained on various question answering (QA) datasets including extractive and multiple-choice QA (Lin et al., 2021; Raffel et al., 2020). Originally, it was proposed for zero-shot dialogue state tracking and utilizes slot description as a question to extract corresponding value from a given input text. Although having descriptions for each data field is not realistic in our case, we manually added the description to each field of the trackers to construct input prompts of TransferQA. For example, "*extractive question: the number of repetitions or laps of the exercise? context: user: i did push-ups for 3 repetitions at light intensity.*" is the input representation to get *Repetitions* of exercise from the case in Figure 1. For the choice and Likert scale fields, the options were included the prompt as well. For GPT-3, we prompted the model to extract field values from an input phrase by giving only the tracker schema without any examples.

In-context Learning We simulated the scenarios where the user provides a phrase when there are zero to four prior items in a database for the corresponding tracker. Since we used synthetic samples as a validation set, we treated one of the sample trackers as a user tracker and excluded the samples for the tracker from the store when augmenting the

---

[1] https://pypi.org/project/sentence-transformers/

[2] https://openai.com/api/

Model	N-shot	JGA	F_1	[a]B-4	[b]R-L
Pure Zero-shot (Baseline)					
TransferQA	0	27.8	53.7	13.6	28.7
GPT-3	0	26.2	49.9	41.2	58.5
In-context Learning (Augmented 10-shot Prompts)					
	0	42.5	68.3	56.3	77.1
	1	51.2	73.1	57.0	78.9
GPT-3	2	**57.7**	75.6	**58.8**	**80.5**
	3	56.9	**76.6**	57.3	78.3
	4	56.5	76.4	57.8	78.8

[a]: BLEU-4, [b]: ROUGE-L

Table 1: Zero and few-shot evaluation results on our validation dataset. The N-shot is the number of examples of the corresponding tracker included in a prompt.

prompt. We iterated over all 24 trackers and 504 items, for each N-shot iteration ($504 \times 5 = 2520$).

Evaluation Metrics   We employed joint goal accuracy (JGA) and F_1 score, which are usually used for the dialogue state tracking tasks to measure the NLU performance of the models. JGA checks whether all predicted values are exactly matched with the ground truth values whereas F_1 checks partial matches between them. For these measures, we excluded *long-form text* fields, for which we instead measured BLEU-4 and ROUGE-L scores.

Results   Table 1 illustrates the evaluation results. In pure zero-shot cases, TransferQA slightly outperformed for close-ended fields (JGA and F_1) but GPT-3 performed almost twice better in extracting open-ended, long-form text fields (B-4 and R-L). In in-context learning cases, GPT-3 augmented with synthetic samples surprisingly outperformed both baseline models in both close-ended and open-ended fields. Even when there were no prior items for the corresponding tracker (zero-shot in in-context learning), JGA was improved by 16.3% and F_1 by 18.4 in GPT-3. The preformance generally increased by the increase of the number of prior items but seemed to be saturated around two- or three-shots.

Limitation   As a preliminary evaluation, we used the synthetic samples as a validation set. For a more ecologically valid evaluation, we need a human-generated dataset in the future.

5   Discussion and Future Directions

In this section, we discuss the rooms for improvement and envision collaborative application domains for both HCI and NLP researchers.

5.1   Prompt Engineering and Seed Sampling

In this work, we mixed the nearest and farthest samples in terms of linguistic similarity between the phrases. A logical next step would be to investigate different strategies to generate prompts. For example, we may hierarchically pick the appropriate trackers first and then retrieve samples from them. Another approach is to split the data fields into groups and run a PLM for each one separately.

5.2   Ethically Boosting Performance through Synthetic Data Augmentation

Self-tracking data is inherently sensitive to privacy issues because they contain personal health and activity history. Therefore, training machine learning models with self-tracking data from multiple people may raise ethical issues (Saltz et al., 2019) and thus is impractical. In contrast, our framework leverages only synthetic samples and the user's own data points to boost up model performance. Our approach demonstrates a feasibility of leveraging common sense of large language models instead of training a baseline model using data collected from a group of people. Future work remains to investigate the external validity of the synthetic samples when the framework serves real-world cases.

5.3   Warm-Starting Self-Tracking in Cold-Start Settings

We note that the in-context learning zero-shot cases in our experiment provide *pure* zero-shot experiences from the users' perspectives; with the framework embedded in a self-tracking tool, the tool is likely to yield the boosted performance even when the user inserts a natural language query for the first time. Going further, since the performance significantly increases with only one or two contextual samples (See Table 1), the user interfaces can be designed to preemptively retrieve a few samples from a new user. For example, the system may nudge the user to provide several example utterances in the initial calibration stage. Designing effective warm-starting interaction warrants further research especially from the HCI perspective.
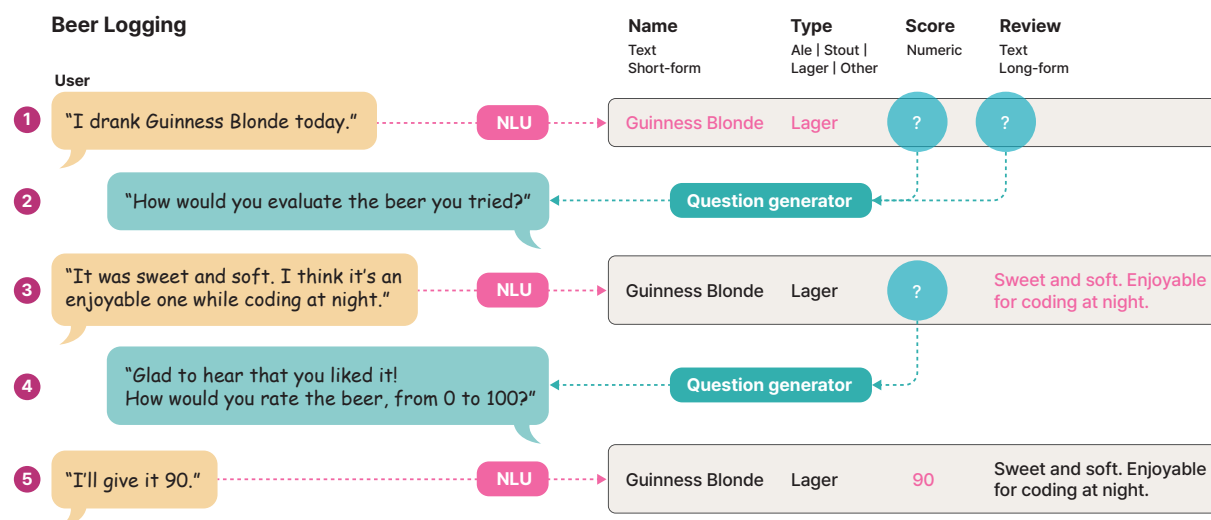
Figure 2: A scenario for a conversational agent embedding our framework (NLU) and a question generator. Upon receiving a user utterance that specifies a subset of field values of the tracker (elements in pink on the right), the question generator suggests a proper response of the system that asks back the user to specify the remaining fields.

5.4 Future Application Domains

Introducing a new NLP task for self-tracking, we propose several application domains to which our approach can be expanded further.

Designing User Interfaces Our topic-agnostic framework can be integrated to a wide range of self-tracking tools in various form factors that support natural language interaction. With speech, the framework can be employed to implement vision- and hands-free tools on smart speakers or smartwatches. Since more than 40% of the trials include erroneous extractions (see JGA in Table 1), proper error recovery methods (*e.g.*, a roll-back button [Kim et al. 2021]) should be provided to users for sustainable interaction.

Multi-Turn Conversation for Data Collection Using a tracker with a long list of data fields, it is not natural to describe all the required field values in a single utterance. Since our framework assumes that the input phrase describes a **subset** of data fields, we can expand the task as a **multi-turn conversation** scenario (*e.g.*, Bae et al. 2022) where the system asks back to fill out missing information in an item. Figure 2 illustrates a scenario of beer logging through a conversational agent, embedding our framework combined with a question generator. This can be also viewed as schema-guided *dialog state tracking* (Rastogi et al., 2020), but the extraction of multiple-choice and long-form text fields poses challenges from the NLP perspective.

Schema-Free Data Collection We are also investigating a more radical scenario where people capture logs even without a predefined tracker schema and the system automatically generates the proper schema based on the natural language phrases. Supporting such schema-free tracking would effectively reduce the learning curve for newcomers to self-tracking tools, especially when the trackers are customizable (Kim et al., 2017).

6 Conclusion

In this work, we introduced a novel NLP task for data collection in self-tracking and presented an NLU framework that can effectively solve this task by augmenting PLM's prompt with synthetic samples. Drawing on the favorable outcomes from the preliminary evaluation, we discussed future research directions regarding improving the pipeline as well as designing user interfaces to effectively support self-tracking through our framework. As an interdisciplinary team of both HCI and NLP researchers, we hope this work inspires other researchers working on the growing areas of self-tracking and personal informatics, where we still need more synergistic collaboration between the NLP and HCI disciplines.
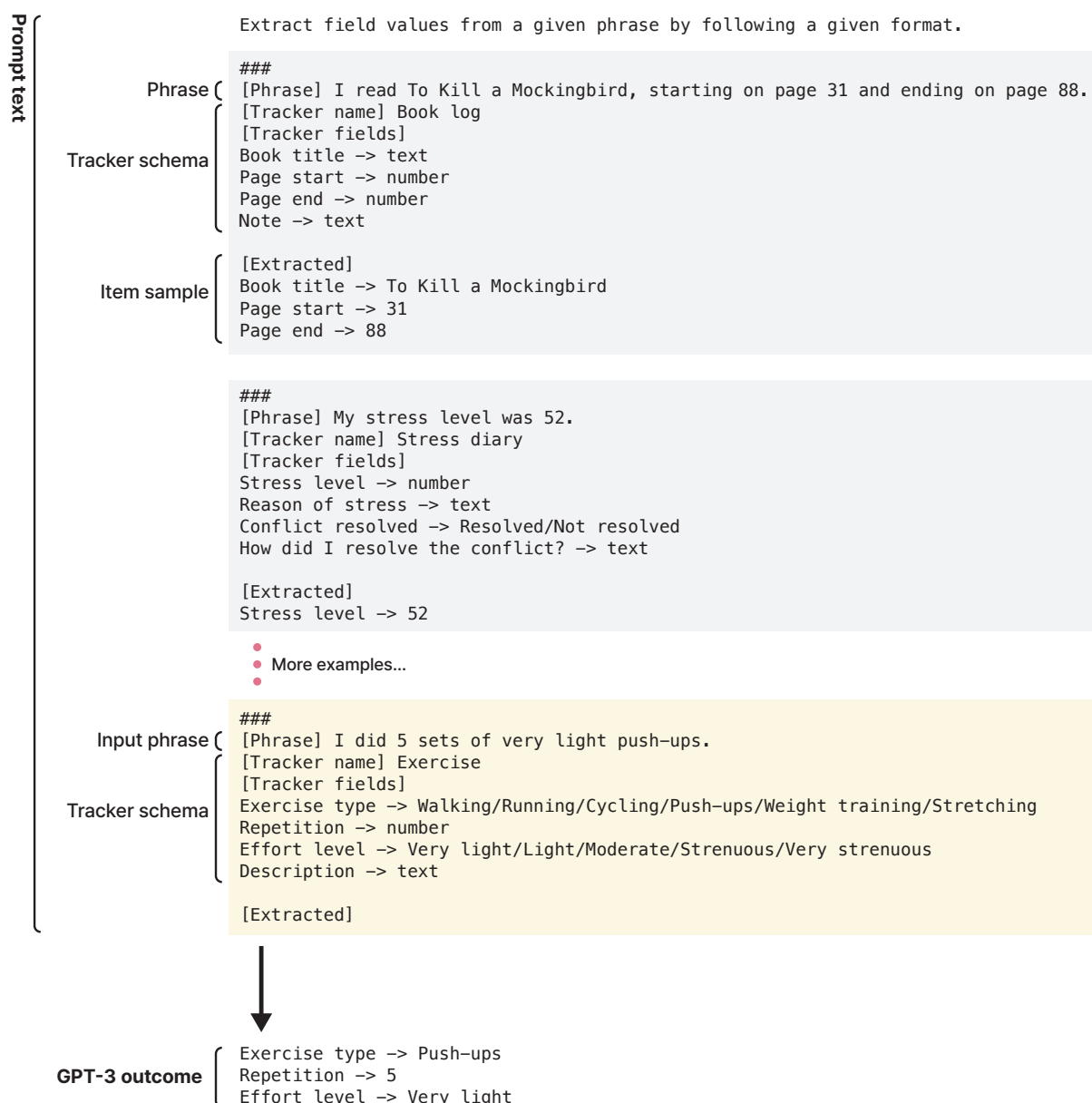
References

Apple Inc. 2021. Health.

Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and

Woomyoung Park. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL '22 (To Appear).

Biowink GmbH. 2021. Clue Period & Ovulation Tracker.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems (NeurIPS '20)*, volume 33, pages 1877–1901. Curran Associates, Inc.

Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding Quantified-selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1143–1152, New York, NY, USA. ACM.

Michael Dietz, Ilhan Aslan, Dominik Schiller, Simon Flutura, Anika Steinert, Robert Klebbe, and Elisabeth André. 2019. Stress Annotations from Older Adults - Exploring the Foundations for Mobile ML-Based Health Assistance. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 149–158, New York, NY, USA. ACM.

Daniel A Epstein, Daniel Avrahami, and Jacob T Biehl. 2016. Taking 5: Work-Breaks, Productivity, and Opportunities for Personal Informatics for Knowledge Workers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 673–684, New York, NY, USA. ACM Press.

Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and taking stock of the personal informatics literature. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4).

Fitbit, Inc. 2021a. Fitbit.

Fitbit, Inc. 2021b. Fitbit Skill.

Jae Ho Jeon. 2016. OmniTrack: Designing Flexible and Highly Customizable Quantified-Self Tool. Ms thesis, Seoul National University, Seoul, Korea.

Yea-Seul Kim, Mira Dontcheva, Eytan Adar, and Jessica Hullman. 2019a. Vocal Shortcuts for Creative Experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14, New York, New York, USA. ACM Press.

Yoojung Kim, Bongshin Lee, and Eun Kyoung Choe. 2019b. Investigating data accessibility of personal health apps. *Journal of the American Medical Informatics Association*, 26(5):412–419.

Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. 2019c. Understanding personal productivity: How knowledge workers define, evaluate, and reflect on their productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 615:1–615:12, New York, NY, USA. ACM.

Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. Mymove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. ACM.

Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):67:1–67:28.

Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. ACM.

Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A Stage-based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 557–566, New York, NY, USA. ACM.

Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul A Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. 2021. Zero-shot dialogue state tracking via cross-task transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7890–7900.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3?

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. Foodscrap: Promoting rich data capture and reflective food journaling through speech input. In *Designing Interactive Systems Conference 2021*, DIS '21, page 606–618, New York, NY, USA. ACM.

Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. Tandemtrack: Shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. ACM.

Lucas M. Silva and Daniel A. Epstein. 2021. Investigating preferred food description practices in digital food journaling. In *Designing Interactive Systems Conference 2021*, DIS '21, page 589–605, New York, NY, USA. Association for Computing Machinery.

MyNetDiary. 2021. MyNetDiary Diabetes.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.*, 19(4).

Under Armour, Inc. 2021. MyFitnessPal Lite (Alexa Skill).

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *The 38th International Conference on Machine Learning (ICML '21)*, pages 12697–12706.

A Exemplary Prompt for GPT-3

```
Extract field values from a given phrase by following a given format.

###
[Phrase] I read To Kill a Mockingbird, starting on page 31 and ending on page 88.
[Tracker name] Book log
[Tracker fields]
Book title -> text
Page start -> number
Page end -> number
Note -> text

[Extracted]
Book title -> To Kill a Mockingbird
Page start -> 31
Page end -> 88

###
[Phrase] My stress level was 52.
[Tracker name] Stress diary
[Tracker fields]
Stress level -> number
Reason of stress -> text
Conflict resolved -> Resolved/Not resolved
How did I resolve the conflict? -> text

[Extracted]
Stress level -> 52
```

More examples...

```
###
[Phrase] I did 5 sets of very light push-ups.
[Tracker name] Exercise
[Tracker fields]
Exercise type -> Walking/Running/Cycling/Push-ups/Weight training/Stretching
Repetition -> number
Effort level -> Very light/Light/Moderate/Strenuous/Very strenuous
Description -> text

[Extracted]
```

Labels on the left: **Prompt text**, *Phrase*, *Tracker schema*, *Item sample*, *Input phrase*, *Tracker schema*

GPT-3 outcome

```
Exercise type -> Push-ups
Repetition -> 5
Effort level -> Very light
```

B Seed Trackers

We manually composed 24 seed trackers (*i.e.*, data schema for tracking). We first extracted 10 tracking themes from a survey of self-tracking research (Epstein et al., 2020) and an empirical study on user-defined self-trackers (Kim et al., 2017). Of the four common types of trackers—*timespamper*, *in-situ experience logger*, *daily summary*, and *archive*—identified by Kim et al. 2017, we composed trackers that fall within either *in-situ experience logger* (A data entry denotes one unit of event or episode) or *daily summary* (A data entry denotes a summarized reflection or information of a day). When designing schemas, we referred to prior self-tracking research prototypes or commercial apps that we have public access to the data format. Table 2 summarizes the format of all seed trackers. Note that all trackers include one time-related field (*e.g.*, Date, Time-point, Time-range), which we omitted in the table for brevity.

Table 2: List of the seed trackers that were used for creating synthetic item samples. The Reference column denotes the existing research or commercial apps that informed the design of the schema.

Type/Name	Data Fields				Reference
Exercise					
In-situ **Exercise**	Exercise type *Choice-single*	Repetition *Number*	Intensity *Scale*	Description *Text*	Kim et al. 2022
Daily **Daily exercise**	Exercise done today *Choice-multiple*	Overall satisfaction *Number*			
	Reflections on today's exercise *Text*				
Sleep					
Daily **Sleep diary**	Sleep quality *Scale*	Memo *Text*			
Medication					
In-situ **Pill intake**	Medication *Choice-single*	Number of pills *Number*	Reason of taking *Text*		
Diabetes					
In-situ **Insulin shots**	Type *Choice-single*	Units *Number*			MyNetDiary 2021
In-situ **Blood sugar**	Glucose level *Number*	Measurement Timing *Choice-single*			MyNetDiary 2021
Food					
In-situ **Meal log**	Meal type *Choice-single*	Menu *Text*	Why I ate this food *Text*	Healthy level *Scale*	Luo et al. 2021
In-situ **Visited restaurant**	Restaurant name *Text*	Type of cuisine *Choice-single*	Menus I tried *Text*	Rating for taste *Number*	
	Rating for location *Number*	Rating for hygiene *Number*	Rating for staff *Number*	Review *Text*	
Daily **Daily eating**	Types of meals had *Choice-multiple*	Relfection on today's eating *Text*			
Beverage					
In-situ **Beverage log**	Category *Choice-single*	Name *Text*	Temperature *Choice-single*	Cups *Number*	Location *Choice-single*
In-situ **Beer log**	Name *Text*	Beer category *Choice-single*	Score *Number*	Review *Text*	Kim et al. 2017
Daily **Daily coffee**	Number of cups had today *Number*	Why I had that amount of coffee *Text*			Luo et al. 2021
Mood					
In-situ **Mood episodes**	Types of mood *Choice-multiple*	Intensity of mood *Number*	Reason of mood *Text*		
In-situ **Stress diary**	Stress level *Number*	Reason of stress *Text*	Conflict resolved *Choice-single*		Dietz et al. 2019
	How did I resolve the conflict? *Text*				

Table 2 – continued from the previous page

Type/Name	Data Fields				Reference
Daily **Daily diary**	Weather Choice-single	Stress level Number	Overall productivity Scale	Atmosphere Choice-single	
	Major types of mood Choice-multiple	Whom I met today Choice-multiple	Reflection on today Text		

Book

In-situ **Book log**	Book title Text	Page start Number	Page end Number	Note Text	

Study

In-situ **Study log**	Study subject Choice-single	Accomplishment Number	Study content Text		
Daily **Study diary**	Study subjects Choice-multiple	Overall satisfaction Number	Reflection on today's study Text		

Productivity

In-situ **Tasks**	Task Choice-multiple	Productivity Scale	Rationale for productivity Text		Kim et al. 2019c
In-situ **Breaks**	What I did during the break Choice-single	Reason for break Text			Epstein et al. 2016
Daily **Work diary**	Major tasks Choice-multiple	Overall productivity Number	Reflections on today's work Text		

Social

In-situ **People**	Who I met Text	Purpose Choice-single	What I did in detail Text		
	Reflections on the interaction Text				

Smoking

In-situ **Smoking log**	Amount Number	Tobacco name Choice-single	Smoking context Choice-single	Smoked with others Choice-single	

Female

Daily **Daily period**	Bleeding Choice-single	Pain Choice-multiple	Emotions Choice-multiple	Sleep Choice-single	Sex Choice-multiple	Biowink GmbH 2021
	Energy Choice-single	Social Choice-single	Reflection on today's menstruation Text			