When LLMs Can't Help: Real-World Evaluation of LLMs in Nutrition

Karen Jia-Hui Li[1,2] **Simone Balloccu**[3] **Ondrej Dusek**[1] **Ehud Reiter**[4]
[1]Charles University, Faculty of Mathematics and Physics, Czechia
[2]Saarland University, Germany [3]TU Darmstadt, Germany
[4]University of Aberdeen, Scotland, UK
Correspondence: li.karen.jh@gmail.com

Abstract

The increasing trust in large language models (LLMs), especially in the form of chatbots, is often undermined by the lack of their extrinsic evaluation. This holds particularly true in nutrition, where randomised controlled trials (RCTs) are the gold standard, and experts demand them for evidence-based deployment. LLMs have shown promising results in this field, but these are limited to intrinsic setups. We address this gap by running the first RCT involving LLMs for nutrition. We augment a rule-based chatbot with two LLM-based features: (1) message rephrasing for conversational variety and engagement, and (2) nutritional counselling through a fine-tuned model. In our seven-week RCT (n=81), we compare chatbot variants with and without LLM integration. We measure effects on dietary outcome, emotional well-being, and engagement. Despite our LLM-based features performing well in intrinsic evaluation, we find that they did not yield consistent benefits in real-world deployment. These results highlight critical gaps between intrinsic evaluations and real-world impact, emphasising the need for interdisciplinary, human-centred approaches.[1]

1 Introduction

Every day, individuals make over 200 food-related decisions (Wansink and Sobal, 2007; van Meer et al., 2016). With sedentary lifestyles becoming increasingly common (Park et al., 2020) and global health issues on the rise (Malik et al., 2013; Rowley et al., 2017), scalable interventions are needed. Digital health technologies via mobile devices offer accessible solutions (Vearrier et al., 2018; Senbekov et al., 2020).

In parallel, advances in fine-tuned language models enabled the generation of human-like responses for many practical applications (Wei et al., 2022; Min et al., 2024). This resulted in a general hype and trust—especially among laypeople and companies—in the potential of this technology (Strange, 2024). This also applies to nutrition: LLMs look promising for tasks like meal recommendation, providing dietary advice, and general domain understanding (Niszczota and Rybicka, 2023; Naja et al., 2024; Tsiantis et al., 2024). Randomised controlled trials (RCTs) are required by domain experts before any real-world deployment (Stolberg et al., 2004; Hariton and Locascio, 2018; Baumel et al., 2019), as they are the foundation of evidence-based medicine and give objective measures of real-world impact. However, past evaluations of LLMs in nutrition are intrinsic only. No evidence has been collected regarding the impact of LLMs in real-world nutrition tasks. This includes sustained diet coaching, where users receive feedback on improving their dietary habits (Vrkatić et al., 2022), or nutritional counselling, where tailored empathetic support helps users address more complex dietary issues (Vasiloglou et al., 2019).

We conduct the first extrinsic evaluation of an LLM-enhanced chatbot for these two tasks. We start from a rule-based chatbot capable of scanning users' food diaries to provide tailored insights. Then, we integrate two LLM-based features: (1) a rephrasing module and (2) a nutritional counselling model. The former varies the base templated responses to make communication more engaging, while the latter provides tailored support, comfort, and suggestions for users' specific dietary concerns. In a seven-week RCT with 81 participants, we compare three groups: a group using the full set of features (insights+rephrasing+counselling), an intermediate group (insights+rephrasing), and a base group using only the rule-based chatbot (insights only). We measure dietary outcomes, emotional well-being, and engagement. Ethics details are presented in Section D.

Based on our results, the "promise" of LLMs

[1]We provide all of our code and results at: https://github.com/saeshyra/diet-chatbot-trial

Proceedings of the 18th International Natural Language Generation Conference, pages 753–779
October 29–November 2, 2025.

in nutrition falls short in the real world: the LLM-based features had little to no effect on any of the measures we consider. Our study provides critical insights into the effectiveness of LLMs in nutrition, the safe deployment of these models, and the interdisciplinary challenges of applying them to sensitive domains.

2 Related Work

Digital health interventions can improve accessibility, cost-effectiveness, and patient-centred care (Greaves et al., 2013; Mitchell and Kan, 2019; Taj et al., 2019). For example, telemedicine allows remote consultations (Barbosa et al., 2021; Totten et al., 2022), while wearable devices enable continuous health monitoring (Izmailova et al., 2018; Natalucci et al., 2023). Mobile health interventions can be highly effective in terms of user adherence (Kamal et al., 2015; Müller et al., 2016; Hoeppner et al., 2017; Lee et al., 2018; Oyebode et al., 2020).

In nutrition, chatbots have emerged as a promising tool to promote healthy eating habits, offering food intake tracking (Graf et al., 2015; Kerr et al., 2016), educational content, and motivational messaging (Fadhil and Villafiorita, 2017; Casas et al., 2018; Maher et al., 2020a). Recent advancements in pre-trained language models can further expand their capabilities in this domain, but not without shortcomings. Recent LLMs can generate meal plans and dietary recommendations based on user needs, but their performance decreases in more complex cases (Niszczota and Rybicka, 2023; Naja et al., 2024; Tsiantis et al., 2024).

Beyond meal guidance and recommendations, AI chatbots are being increasingly explored for their potential to provide empathetic support towards behaviour change. Negative emotions can lead to poorer nutritional choices (Devonport et al., 2019; González et al., 2022) and tailored advice can mitigate this mental load (Balloccu and Reiter, 2022a; Park et al., 2024). Chatbots can be approachable, calming, and display adequate therapeutic skills (Zhang et al., 2020; Beilharz et al., 2021; Vowels et al., 2024). This can enhance engagement in sensitive contexts like body image, eating disorders, and relationship counselling. Artificial empathy and personalised interactions from chatbots (Stephens et al., 2019; Rahmanti et al., 2022) can help users in pursuing healthier habits. The therapeutic promise of chatbots also raises important ethical questions. There is still need for deeper integration of empathy, mental health sensitivity, and patient safety into chatbot design (Stein and Brooks, 2017; Anisha et al., 2024). When evaluated by experts, AI nutritional support is often scientifically sound but potentially outdated, inaccurate (Kirk et al., 2023), or even harmful for more complex cases (Balloccu et al., 2024).

While accuracy is important, user engagement also plays a critical role to the success of digital health interventions in nutrition. When users lose interest in using the chatbots, this causes a rapid decrease in dietary adherence, and eventually result in early drop-out (Fadhil, 2018; Maher et al., 2020b; Balloccu et al., 2021). User satisfaction and motivation are usually pursued through personalisation, communication, visual elements (Kettle and Lee, 2024; Balloccu and Reiter, 2022b), gamification (Fadhil and Villafiorita, 2017), or social support mechanisms (Svetkey et al., 2015).

For all of the above aspects, rigorous extrinsic evaluation is needed, to move chatbots from experimental prototypes to deployable healthcare assistants (Baumel et al., 2019). Yet, existing evaluations are synthetic (Mishra et al., 2024; Yang et al., 2024), focused on textual characteristics or accuracy against standardised benchmarks (Parameswaran et al., 2024; Azimi et al., 2025). Our work is, to our knowledge, the first (Omar et al., 2024) randomised controlled evaluation of LLM-delivered diet coaching and nutritional counselling over an extended deployment.

3 Chatbot Development

We extend the rule-based diet-coaching chatbot by Balloccu and Reiter (2022b), which is deployed on the Telegram messaging platform and designed to deliver personalised nutritional insights based on users' food diary on MyFitnessPal (Evans, 2017), a popular and freely available calorie-counting app. The core functionality of the chatbot involves delivering insights in two forms: (1) *basic insights* (Figure 2a), which are simple recaps of a user's dietary intake—calories and nutrients, and (2) *advanced insights* (Figure 2b), which present an extended textual description with some corresponding data visualisation.

For this work, we extend the chatbot[2] with two LLM-powered features: rephrasing to vary the templated responses, and nutritional counselling through fine-tuned models. Figure 1 illustrates

[2]We use the code made available by the authors.

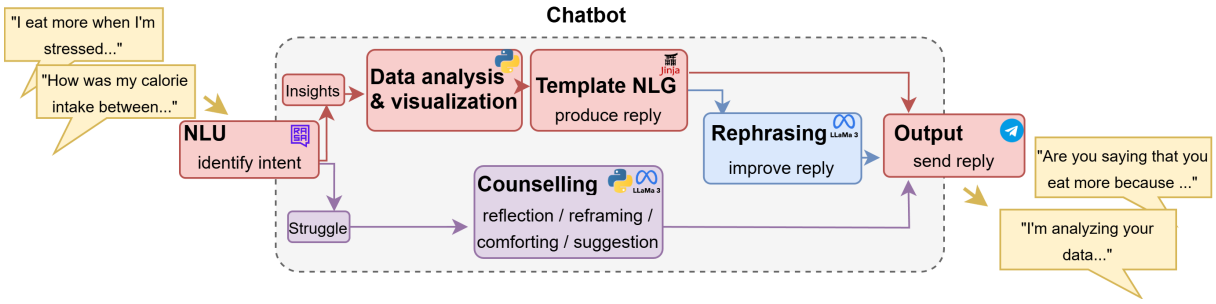
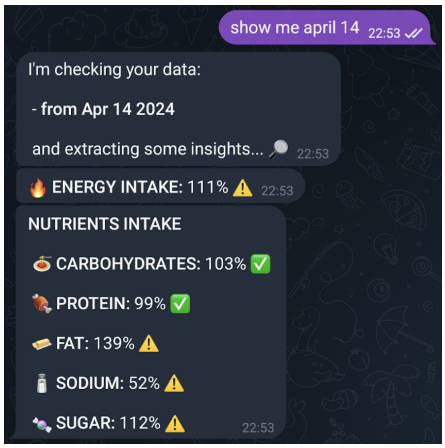
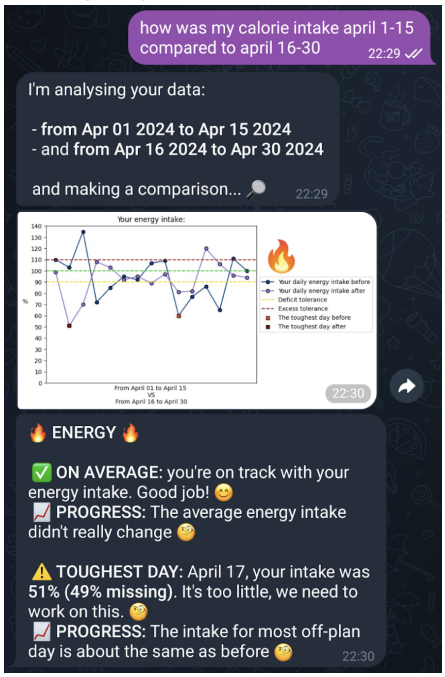


Figure 1: Overview of the chatbot architecture and functional flow. The BASELINE version uses the red flow only, REPHRASED adds the step marked in blue, and FULL adds the flow marked in purple. We provide an example of the insights flow in Figure 2 and the supportive text flow in Table 2.



(a) **Basic insights** into all monitored nutrients for a single day.



(b) **Advanced insights** comparing the average calorie intake and toughest day from the first half of the month to the other half.

Figure 2: Examples of the chatbot outputs.

the architecture and flow of our chatbot. Its main objective is to improve long-term diet adherence, emotional well-being, and user engagement.

3.1 Rephrased Responses

The original chatbot code included a small set of slots used to vary the templated responses. This system ensured consistency and safety, but lacked the conversational fluidity of human dialogue. To enhance communication variety, we prompt an LLM to rephrase the templated outputs, while maintaining clarity for more structured messages. These enhancements aim to enable more varied communication with the chatbot, and encourage higher engagement among trial participants.

To achieve high-quality rephrasing without the need for additional fine-tuning, we experimented with prompt engineering using instruction-tuned variants of Gemma 7B (Gemma Team et al., 2024), Mistral 7B (Jiang et al., 2023), and Llama 3 8B (AI@Meta, 2024), settling on Llama 3 8B as the production model.

With basic prompting (as in Figure 3), we faced issues with ambiguous and context-sensitive message templates (Figure 4). To address these, we exploit the fact that the rule-based chatbot gives us explicit access to the output message intent (e.g. insights on which nutrients, over which days). We develop a targeted prompt that dynamically adapts to message context with explicit, intent-specific instructions (Figure 5), reducing hallucinations by constraining the model to rephrase within context.

3.1.1 Evaluation

We conducted an additional human evaluation with 20 native English-speaking crowd workers on Prolific. Participants were shown pairs of templated and rephrased messages and asked to indicate which they preferred, which felt more natural, and

Rephrase the following message to the user, keeping any mentioned dates. Do not introduce new dates or assume time periods. Do not add extra information. Use emojis.
`[message]`

Figure 3: Initial prompt for message rephrasing.

Example Templated Output

⚠️ **TREND AND CONSISTENCY:** sorry, I need 3 or more days for this 😕

Example Rephrased Output with Initial Prompt

⚠️ Need a little time to get into the swing of things! 😌 Please allow at least 3 days for this ⏰

Figure 4: Example of a problematic rephrased output from the initial rephrasing prompt (Figure 3), due to ambiguity in the original templated message responding to a user's request for advanced insights over a time period shorter than three days.

Final Rephrasing Prompt
`INTENT: "compare_no_dates"; NUTRIENT: None;`

The user requested comparative insights into their food diary but did not give dates. Do not greet the user. Do not include additional information. Use simple language and emojis. Rephrase the following message to the user.

Rephrase the message:
Please give me two dates or a date range to compare

Example Rephrased Output

To help you with your food diary, could you please provide two specific dates or a date range for comparison? 🗓️👀

Figure 5: An example of the dynamic rephrasing. The context leading up to the intent of the templated chatbot output ("`compare_no_dates`" + no nutrient specified) is extracted from the NLU pipeline and dynamically added to the prompt, resulting in the rephrased output.

whether both conveyed the same meaning. About 65% of responses were preferred in their rephrased form, and 72% were judged more natural. Only a small number of participants (n=6) reported any differences in meaning between the messages. These findings confirm that LLM-based rephrasing, when carefully prompted, can enhance the linguistic quality and engagement of chatbot responses. We provide more information on this evaluation in Section C.

3.2 Nutritional Counselling

We integrate a nutritional counselling feature designed to support users not only with dietary data, but also with the psychological and behavioural challenges of healthy eating. This feature is powered by a language model fine-tuned on the `HAI-Coaching` dataset (Balloccu et al., 2024), a collection of ∼2.4K crowd-sourced dietary *struggles* paired with ∼100K expert-annotated supportive responses. The *struggles* are textual descriptions of problems affecting people's diet and cover a wide variety of topics, from snacking and dietary restrictions, to emotional eating, anxiety, and depression. The responses are equally split into four categories—*reflection* (understanding the struggle), *comfort* (providing emotional support), *reframing* (portraying the struggle positively), and *suggestion* (actionable next steps)—as curated by human experts and reflective of the psychological research behind nutritional counselling.

We initially test by fine-tuning several models: GPT-2 medium (Radford et al., 2019), FLAN-T5 base (Chung et al., 2024), BabyLlama (Timiryasov and Tastet, 2023), Gemma 7B (Gemma Team et al., 2024), Mistral 7B (Jiang et al., 2023), and Llama 3 8B (AI@Meta, 2024). We include older and more limited models (GPT-2 and BabyLlama) to inspect whether newer, instruction-tuned ones offer a real advantage in terms of performance.

For GPT-2 and BabyLlama, we guide the generation of each category of supportive text via special tokens ("`<|struggle|>`, "`<|reflection|>`, etc.) in a controllable text generation fashion (Keskar et al., 2019; Li and Liang, 2021; Zhang et al., 2023). For instruction-tuned models, we use category-specific prompts, mirroring those used to create the dataset (Balloccu et al., 2024). We provide the prompts in Table B.1.

Following common practices in NLG intrinsic evaluation (Sai et al., 2022), we calculate BLEU-1/3 and BLEURT (Table 1), and also conduct a qualitative review of the generated outputs. We immediately excluded GPT-2, BabyLlama, Gemma, and Mistral, as they showed poor performance and consistently failed in producing useful supportive text. FLAN-T5 showed strong BLEU-1 performance, but frequently produced contradictory or irrelevant responses. Llama 3, on the other hand, delivered semantically coherent and contextually appropriate suggestions across a range of examples,

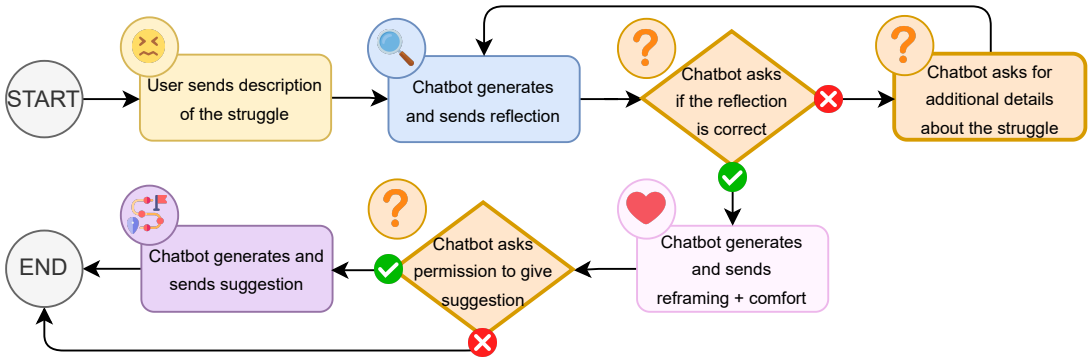


Figure 6: Nutritional counselling flow. When receiving a dietary struggle, the chatbot generates and sends reflections, asking the user for feedback (up to a limited amount of retries). Once a correct reflection is generated, comfort and reframing statements are generated and sent. Following this, the chatbot asks for permission to send a suggestion, ending the conversation.

Model	BLEU-1	BLEU-3	BLEURT
GPT-2 medium	10.70	22.45	-0.38
BabyLlama	10.30	22.27	-0.42
FLAN-T5 base	**20.31**	34.78	-0.16
Gemma 7B	17.51	27.06	-0.20
Llama 3 8B	19.16	**36.48**	**-0.14**
Mistral 7B	12.50	25.47	-0.18

Table 1: Results from automatic evaluation of fine-tuned nutritional counselling models, using BLEU with 1 reference, BLEU with 3 references, and BLEURT. The best value for each metric is displayed in **bold**.

Struggle: I eat more because of stress
Reflection: Are you saying that you eat more when you are feeling stressed?
Comfort: It's understandable to turn to food as a coping mechanism for stress.
Reframing: Something positive you could consider about this is that you are aware of the connection between stress and eating habits, and can take steps to address it.
Suggestion: Starting from tomorrow, you could try practicing mindfulness techniques such as deep breathing or meditation to help manage stress and reduce cravings for food.

Table 2: Example outputs of each supportive text category from the fine-tuned Llama 3 model, in response to a user-given struggle.

achieving the best BLEU-3 and BLEURT scores. Based on these findings, we selected Llama 3 for integration into the final chatbot.

We provide an example output from Llama 3, fine-tuned for nutritional counselling, in Table 2. Finally, we implement a conversation flow to guide users seeking help, as outlined in Figure 6.

4 Chatbot Trial

We conducted a seven-week randomised controlled trial to test the chatbot's real-world impact from May to June 2024. Participants were recruited through social media, flyers, and direct outreach across multiple locations, resulting in a demographically diverse cohort. Exact demographics are outlined in Section F. Through an onboarding process, users were taught how to install the required apps, connect their accounts, access the chatbot, and use it. We successfully onboarded 87 participants, and 81 of them completed the full seven-week duration. Dropouts occurred across all study groups, primarily due to time constraints or lack of engagement. These early exits were balanced across conditions and did not compromise the integrity of the trial.

Participants were randomly assigned to one of three groups: BASELINE (n=26), with templated insights from the food diary only; REPHRASED (n=27), with LLM-rephrased responses; FULL (n=28), with LLM-rephrased messages and nutritional counselling. Throughout the trial, participants logged their meals daily via MyFitnessPal, engaged with the chatbot on Telegram five or more times per week, and completed a weekly emotional well-being questionnaire using the Positive and Negative Affect Schedule (PANAS) (Thompson, 2007) (more details in Section 5.2). Participation in the trial was incentivised through weekly online gift vouchers, with a doubled reward in the final week, and adherence was monitored through MyFitnessPal logs and conversation history. While we en-

Group	kcal (%)			Carbs (%)			Protein (%)			Fat (%)			Sodium (%)			Sugar (%)		
	$W1(\downarrow)$	$W7(\downarrow)$	$\Delta(\uparrow)$	$W1(\downarrow)$	$W7(\downarrow)$	$\Delta(\uparrow)$	$W1(\downarrow)$	$W7(\downarrow)$	$\Delta(\uparrow)$	$W1(\downarrow)$	$W7(\downarrow)$	$\Delta(\uparrow)$	$W1(\downarrow)$	$W7(\downarrow)$	$\Delta(\uparrow)$	$W1(\downarrow)$	$W7(\downarrow)$	$\Delta(\uparrow)$
BASELINE	**21.96**	**21.32**	**0.64**	34.95	32.50	2.45	**35.41**	**30.87**	**4.54**	38.97	38.49	**0.48**	**50.73**	49.64	1.09	52.93	50.34	**2.58**
REPHRASED	25.42	24.33	1.09	**31.89**	32.80	**-0.91**	35.75	34.16	1.58	**35.67**	**37.11**	**-1.44**	51.96	61.07	**-9.11**	**49.03**	**49.34**	-0.31
FULL	26.68	23.40	**3.28**	37.56	**31.22**	**6.34**	36.04	34.63	**1.42**	36.21	37.17	-0.97	57.78	50.28	**7.49**	49.95	53.11	**-3.16**

Table 3: Group adherence to dietary goals. We report the absolute distance (%) from calories and nutrient goals (on average per group). We compare differences (Δ) in average between the first ($W1$) and last ($W7$) week of trial. **Best** and **worst** Δ highlighted.

Group Comparison	kcal		Carbs		Fat		Protein		Sodium		Sugar	
	Diff.	p-value	Diff.	p-value	Diff.	p-value	Diff.	p-value	Diff.	p-value	Diff.	p-value
BASELINE - REPHRASED	-0.27	0.54	0.65	0.27	-0.16	0.84	-0.15	0.86	0.14	0.87	0.22	0.81
BASELINE - FULL	-0.38	0.38	-0.69	0.23	0.17	0.84	-0.31	0.71	-0.88	0.29	0.02	0.98
REPHRASED - FULL	-0.11	0.79	-1.34	**0.02***	0.33	0.68	-0.16	0.85	-1.02	0.21	-0.20	0.82

Table 4: Differences and p-values from the mixed-effects models comparing group pairs for energy and nutrients goals. We compare weekly changes per-metric for each group. Significant p-values are marked with an asterisk (*).

couraged regular participation, occasional lapses were tolerated as long as participants remained responsive and completed the weekly questionnaires. Ethics and exact compensation details are provided in Section D.

During the trial, participant adherence was monitored using automated checks, including morning checks of the previous day's food log to identify incomplete or implausible entries, mid-afternoon reminders for missing entries, and evening nudges for inactive users (more details in Section A). While nutritional counselling could be accessed anytime by the FULL group, the chatbot also actively offered it each Friday. At the end of each week, the chatbot provided participants with a link to the PANAS questionnaire to assess emotional well-being and released their weekly voucher upon completion. At the conclusion of the trial, offboarding had participants fill out a final feedback form tailored to their assigned study group.

The trial faced several technical challenges, including a temporary disruption in the chatbot's access to MyFitnessPal data (beyond our control), a necessity to retrain the nutritional counselling model, and rare bugs that made the chatbot unusable for short periods of time (typically one hour or less, and fixed promptly). We discuss these in Section E.

5 Results

5.1 Dietary outcome

Participants in each group logged their daily dietary intake using MyFitnessPal, allowing us to evaluate adherence to the personal diet goals provided by the app. We focused on how the LLM-powered features in REPHRASED and FULL influenced intake behaviours compared to BASELINE. We first measure the absolute distance (%) from user's intake goals (lower is better). Since different groups started at different distances, we consider the difference between the first and last weeks of the trial as a more objective measure. Using this approach, we avoid cases in which an improvement could be observed simply because the relevant group started, on average, closer to a specific goal.

An initial look at results (Table 3) looks promising: for the three metrics where FULL shows the greatest improvement (kcal, carbs and sodium), the values proved 2x-7x more than that of BASELINE. REPHRASED, on the other hand, never showed a greater improvement than the other groups and actually showed the worst result out of the three metrics.

However, the improvement values fall within a very small range: we see cases where the "best" group shows less than a 1% improvement, and even the greatest values do not go past ~7.5%. Therefore, we check for any significance in these results through a linear mixed-effects model (Table 4). Here, the narrative changes: we find no significance except for a group-by-time interaction for carbohydrate adherence in FULL compared to REPHRASED, hinting at an improved alignment with carbohydrate targets over time. Considering the lack of significance for any other measure, we deem this to be insufficient evidence of the benefits provided by LLM-based features. We report further insights and visualisation on dietary outcomes in Section G.

Group	PA (↑)			NA (↓)		
	$W1$	$W7$	$\Delta(\uparrow)$	$W1$	$W7$	$\Delta(\downarrow)$
BASELINE	**15.52**	15.40	**-0.12**	8.56	**8.44**	-0.12
REPHRASED	14.81	**16.77**	**1.96**	10.69	9.77	**-0.92**
FULL	14.30	14.30	0.00	**8.26**	8.78	**0.52**

Table 5: Average positive affect (PA) and negative affect (NA) scores in week 1 and week 7, and the delta between them. We compare difference (Δ) in average between the last ($W7$) and first ($W1$) week of trial. **Best** and **worst** Δ highlighted.

Group	PA		NA	
	Diff.	**p-value**	**Diff.**	**p-value**
BASELINE - REPHRASED	0.14	0.34	-0.13	0.47
BASELINE - FULL	-0.05	0.73	0.18	0.31
REPHRASED - FULL	-0.19	0.20	0.31	0.08

Table 6: Differences and p-values from the mixed-effects models for the PANAS scores. We compare weekly changes per-score for each group. Significant p-values are marked with an asterisk (*).

5.2 Emotional Well-being

The PANAS questionnaire is a validated self-assessment tool to measure emotional affect. Participants are asked to rate specific emotions on a scale of one to five, based on the extent they felt them in the past week. From this, PANAS returns two independent scores: positive affect (PA, higher is better) and negative affect (NA, lower is better). To analyse the effect of LLM-based features on emotional state, we monitor the weekly change in both PA and NA. Ideally, we would expect a noticeable improvement in FULL, since this group had access to the nutritional counselling feature, allowing them to receive tailored empathetic support.

The overall results (Table 5) do not show particularly evident trends. FULL had no change in PA, and a negligible worsening in NA. In contrast, REPHRASED displayed the largest PA and NA improvements. BASELINE exhibited opposite trends for the two measures, with a slight decline in PA but a similarly small improvement in NA.

Again, the observed changes are relatively small. The fact that the participants from REPHRASED demonstrated the greatest improvements to emotional wellbeing out of all groups contradicts our hypotheses. This pattern would suggest that rephrasing alone noticeably boosts emotional wellbeing, while nutritional counselling has the opposite effect. As before, we run a linear mixed-effects model Table 6 to inspect whether any of these changes are statistically significant. The model did not identify any significant effects, including for REPHRASED. We further analyse the individual emotions targeted by PANAS by breaking down the scores (more information in Section H). However, no emerging trend or significance was found.

5.3 User Engagement

To investigate changes in engagement, we calculated the number of interactions (individual messages from the user), conversations (a sequence of interactions with responses within five minutes of each other), and days of chatbot use over the seven-week intervention. We report more detailed engagement metrics in Section I. The utility of LLM-based functions here cause a noticeable increase in the above metrics, indicating that the FULL and REPHRASED groups spent more time interacting with the chatbot.

Our results (Table 7) show a consistent decline in all engagement metrics over time, regardless of the group. This indicates a natural decrease in user engagement over the weeks. FULL consistently showed higher interaction levels, likely driven by the additional nutritional counselling functionality and the weekly direct prompt encouraging feature use. In contrast, REPHRASED did not exhibit notable differences in engagement compared to BASELINE. Using a linear mixed-effects model, we find only one significant difference: participants in FULL spent significantly more days interacting with the chatbot than those in BASELINE. This suggests that people in FULL spent significantly more days interacting with the chatbot. Given the additional promotion of the "advice" feature in FULL, the lack of significant differences across other metrics, and the overall downward trend, we do not consider this sufficient evidence to support consistent benefits of LLM-based features.

5.4 User Feedback

User feedback collected at the end of the trial provided insight into the perceived strengths and weaknesses of the chatbot. Over 39% of participants from all groups judged the visualisations accompanying the advanced insights as helpful for understanding their nutritional data. Users also appreciated the food diary reminders, time-period comparisons, nutritional breakdowns and the chatbot's conversational tone and ease of use.

Group	Interactions			Conversations			Days		
	$W1(\uparrow)$	$W7(\uparrow)$	$\Delta(\uparrow)$	$W1(\uparrow)$	$W7(\uparrow)$	$\Delta(\uparrow)$	$W1(\uparrow)$	$W7(\uparrow)$	$\Delta(\uparrow)$
BASELINE	23.15	9.77	**-13.38**	**8.08**	5.19	**-2.88**	**5.31**	4.38	**-0.92**
REPHRASED	24.30	10.22	-14.07	7.74	5.63	-2.11	5.30	4.74	-0.56
FULL	**27.75**	**13.39**	**-14.36**	7.54	**6.25**	**-1.29**	4.93	**5.14**	**0.21**

Table 7: The count of interactions, conversations (interactions with less than 5 minutes in between), and interaction days across each group. We compare difference (Δ) in average between the first ($W1$) and last ($W7$) week of trial. **Best** and **worst** Δ highlighted.

Group Comparison	Interactions		Conversations		Days	
	Diff.	p-value	Diff.	p-value	Diff.	p-value
BASELINE - REPHRASED	-0.02	0.96	0.08	0.53	0.07	0.26
BASELINE - FULL	-0.22	0.67	0.21	0.11	0.14	**0.02***
REPHRASED - FULL	-0.20	0.69	0.13	0.34	0.07	0.25

Table 8: Differences in engagement metrics (by count) and p-values from the mixed-effects models. Significant p-values are marked with an asterisk (*).

However, many participants (50% of FULL, 59% of REPHRASED, and 32% of BASELINE) reported difficulties with the chatbot's NLU module, specifically when typos were present or phrasing deviated from expected patterns. Some also struggled with fully using the chatbot despite the manual provided. An additional concern was the chatbot's limited functionality when compared to tools like ChatGPT. This reflects users' increasing expectations shaped by open-ended LLMs, although in healthcare contexts, stricter safety and reliability constraints apply. Although hallucinations were rarely reported, several users were frustrated by vague insights or inaccuracies stemming from the food diary data, and requested more tailored and concrete suggestions. In particular, users wanted meal recommendations, which we had to exclude because of safety compliance.

Regarding nutritional counselling, some participants appreciated the supportive tone, while others criticised the advice as overly generic, unhelpful, or even counterproductive. Some users felt that the attempt to offer comfort detracted from the clarity and speed of getting useful responses. One participant specifically criticised the open-ended nature of the advice, noting that statements like, "You could consider that snacking can be a way to fuel your body and give it the energy it needs," risk inadvertently endorsing unhealthy habits. These issues align with the results reported by the dataset authors (Balloccu et al., 2024).

6 Conclusion

As of today, there is an increasing trust in applying LLMs to healthcare and its related sub-domains, including nutrition. However, these models are typically evaluated intrinsically, on unrealistic or simulated tasks, and mostly relying on metrics. In this work, we present the first objective evaluation of the real-world impact of LLMs in nutrition. We integrated them into a chatbot for diet coaching and nutritional counselling, and ran a seven-week RCT on a population of 81 participants.

Our results quickly point out the limits of intrinsic LLM evaluation. We found no consistent improvement across our metrics. Although some statistically significant improvements emerged, these appear spurious in the larger context of our trial. Based on our results, these models are unable to effectively promote dietary adherence, reduce the emotional load of dieting, provide empathetic help through counselling, or simply boost user engagement.

We conclude that, at the current time, there is no evident benefit in applying LLMs to nutrition in the setup we investigated. While our findings are specific to this domain and trial configuration, they serve as a step towards real-world evaluation in healthcare, highlighting how LLMs might (or might not) affect user outcomes in chatbot-driven nutritional counselling. Our overall research underscores the importance of critical evaluation of LLMs in health-focused applications. Although these models are often heralded for their potential to deliver dynamic and personalised interactions,

our findings caution against adoption without rigorous real-world validation. We hope our results will shed light on the need for real-world assessments beyond benchmarks.

7 Limitations

We faced several limitations in this work that highlight areas for improvement in the development and deployment of the diet-coaching chatbot. One significant issue was the natural language understanding component of the chatbot. Despite training Rasa's pipeline on a sufficient number of in-domain examples, the chatbot struggled with the varied user inputs, particularly in processing different date formats, which the chatbot relied on an entity parser to extract. Users often requested insights in diverse ways and with various date notations, leading to occasional misunderstandings or failures to deliver the requested insights. This challenge impacted the chatbot's interactions, and the overall user experience.

Furthermore, the trial's reliance on a task-specific chatbot exposes its limited adaptability when compared to open-domain models that the general public is usually exposed to, such as ChatGPT. Unfortunately, in sensitive domains, task-specific design is mandatory, as it allows more focused and controlled interventions. An RCT with an open-ended chatbot, able to answer any query from the user would have been too dangerous to be allowed. Because of this, we had to restrict our chatbot's ability to handle diverse conversational contexts or unexpected queries, which are strengths of modern open-domain models. This highlights a trade-off between specialisation and adaptability.

Another limitation stemmed from the dataset used to train the nutritional counselling feature. `HAI-Coaching`, while expert-annotated as safe, also contains overly generic advice. Consequently, the chatbot's recommendations often lacked the depth and personalisation necessary for more impactful dietary advice. This issue was further aggravated by the lack of integration between user-reported struggles and their submitted food diaries. Although the food diary data were collected and processed separately for analysis of the trial outcomes, the chatbot did not use them to tailor its advice. This omission, which several trial participants pointed out, could further enhance the nutritional counselling feature.

Additionally, our implementation relied on relatively smaller models rather than larger, more powerful models. This choice was primarily driven by hardware limitations and the practical necessity for fast inference times; the chatbot needed to handle concurrent interactions with all trial participants and provide responses within a few seconds to maintain conversational flow. It remains possible that larger models, with greater capacity and more nuanced language understanding, might have delivered improved counselling quality or engagement.

Finally, as this is the first RCT evaluating LLM-based nutritional counselling, there is limited prior evidence on the timescale over which such interventions might influence dietary outcomes. Unlike previous trials using template-based messages, there are no directly comparable studies to guide expected effect onset. Consequently, the seven-week duration of our trial was chosen pragmatically, constrained by available funding and resources rather than established guidance. Future work could investigate optimal intervention duration, potentially informed by emerging evidence or longitudinal studies, to better contextualise the timing and magnitude of effects.

8 Acknowledgements

This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project (contract no. 812882) and the European Research Council (Grant agreement No. 101039303 NG-NLG). It is further supported by the DYNAMIC center, funded by the LOEWE program of the Hessian Ministry of Science and Arts (grant number: LOEWE1/16/519/03/09.001(0009)/98).

References

AI@Meta. 2024. Llama 3 model card.

Sadia Azmin Anisha, Arkendu Sen, and Chris Bain. 2024. Evaluating the potential and pitfalls of ai-powered conversational agents as humanlike virtual health carers in the remote management of noncommunicable diseases: Scoping review. *Journal of Medical Internet Research*, 26:e56114.

Iman Azimi, Mohan Qi, Li Wang, Amir M Rahmani, and Youlin Li. 2025. Evaluation of llms accuracy and consistency in the registered dietitian exam through prompt engineering and knowledge retrieval. *Scientific Reports*, 15(1):1506.

Simone Balloccu and Ehud Reiter. 2022a. Beyond calories: evaluating how tailored communication reduces emotional load in diet-coaching.

Simone Balloccu and Ehud Reiter. 2022b. Comparing informativeness of an nlg chatbot vs graphical app in diet-information domain. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 156–185.

Simone Balloccu, Ehud Reiter, Matteo G Collu, Federico Sanna, Manuela Sanguinetti, and Maurizio Atzori. 2021. Unaddressed challenges in persuasive dieting chatbots. In *Adjunct Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, pages 392–395.

Simone Balloccu, Ehud Reiter, Karen Jia-Hui Li, Rafael Sargsyan, Vivek Kumar, Diego Reforgiato, Daniele Riboni, and Ondrej Dusek. 2024. Ask the experts: sourcing a high-quality nutrition counseling dataset through human-AI collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11519–11545, Miami, Florida, USA. Association for Computational Linguistics.

William Barbosa, Kina Zhou, Emma Waddell, Taylor Myers, and E. Ray Dorsey. 2021. Improving access to care: Telemedicine across medical domains. *Annual Review of Public Health*, 42:463–481.

Amit Baumel, Frederick Muench, Stav Edan, and John M Kane. 2019. Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *Journal of Medical Internet Research*, 21:e14567.

Francesca Beilharz, Suku Sukunesan, Susan L. Rossell, Jayashri Kulkarni, and Gemma Sharp. 2021. Development of a positive body image chatbot (kit) with young people and parents/carers: Qualitative focus group study. *Journal of Medical Internet Research*, 23.

Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2018. Food diary coaching chatbot. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1676–1680. ACM.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Tracey J Devonport, Wendy Nicholls, and Christopher Fullerton. 2019. A systematic review of the association between emotions and eating behaviour in normal and overweight adult populations. *Journal of Health Psychology*, 24:3–24.

Daniel Evans. 2017. Myfitnesspal. *British Journal of Sports Medicine*, 51(14):1101–1102.

Ahmed Fadhil. 2018. Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation. *arXiv preprint arXiv:1802.09100*.

Ahmed Fadhil and Adolfo Villafiorita. 2017. An adaptive learning with gamification & conversational uis. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 408–412. ACM.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Cristina Elizabeth Fuente González, Jorge Luis Chávez-Servín, Karina de la Torre-Carbot, Dolores Ronquillo González, María de los Ángeles Aguilera Barreiro, and Laura Regina Ojeda Navarro. 2022. Relationship between emotional eating, consumption of hyperpalatable energy-dense foods, and indicators of nutritional status: A systematic review. *Journal of Obesity*, 2022:1–11.

Bettina Graf, Maike Krüger, Felix Müller, Alexander Ruhland, and Andrea Zech. 2015. Nombot – simplify food tracking. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multi-*

media, volume 30-November-2015, pages 360–363. ACM.

Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. 2013. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Quality & Safety*, 22:251–255.

Eduardo Hariton and Joseph J Locascio. 2018. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG : an international journal of obstetrics and gynaecology*, 125:1716.

Bettina B. Hoeppner, Susanne S. Hoeppner, and Lorien C. Abroms. 2017. How do text-messaging smoking cessation interventions confer benefit? a multiple mediation analysis of text2quit. *Addiction*, 112:673–682.

Elena S. Izmailova, John A. Wagner, and Eric D. Perakslis. 2018. Wearable devices in clinical trials: Hype and hypothesis. *Clinical Pharmacology and Therapeutics*, 104:42–52.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ayeesha Kamran Kamal, Quratulain Shaikh, Omrana Pasha, Iqbal Azam, Muhammad Islam, Adeel Ali Memon, Hasan Rehman, Masood Ahmed Akram, Muhammad Affan, Sumaira Nazir, Salman Aziz, Muhammad Jan, Anita Andani, Abdul Muqeet, Bilal Ahmed, and Shariq Khoja. 2015. A randomized controlled behavioral intervention trial to improve medication adherence in adult stroke patients with prescription tailored short messaging service (sms)-sms4stroke study. *BMC Neurology*, 15:212.

Deborah A. Kerr, Amelia J. Harray, Christina M. Pollard, Satvinder S. Dhaliwal, Edward J. Delp, Peter A. Howat, Mark R. Pickering, Ziad Ahmad, Xingqiong Meng, Iain S. Pratt, Janine L. Wright, Katherine R. Kerr, and Carol J. Boushey. 2016. The connecting health and technology study: a 6-month randomized controlled trial to improve nutrition behaviours using a mobile food record and text messaging support in young adults. *International Journal of Behavioral Nutrition and Physical Activity*, 13:52.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Liam Kettle and Yi Ching Lee. 2024. User experiences of well-being chatbots. *Human Factors*, 66:1703–1723.

Daniel Kirk, Elise van Eijnatten, and Guido Camps. 2023. Comparison of answers between chatgpt and human dieticians to common nutrition questions. *Journal of Nutrition and Metabolism*, 2023:1–9.

Mikyung Lee, Hyeonkyeong Lee, Youlim Kim, Junghee Kim, Mikyeong Cho, Jaeun Jang, and Hyoeun Jang. 2018. Mobile app-based health promotion programs: A systematic review of the literature. *International Journal of Environmental Research and Public Health*, 15:2838.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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)*. Association for Computational Linguistics.

Carol Ann Maher, Courtney Rose Davis, Rachel Grace Curtis, Camille Elizabeth Short, and Karen Joy Murphy. 2020a. A physical activity and diet program delivered by artificially intelligent virtual health coach: Proof-of-concept study. *JMIR mHealth and uHealth*, 8:e17558.

Carol Ann Maher, Courtney Rose Davis, Rachel Grace Curtis, Camille Elizabeth Short, and Karen Joy Murphy. 2020b. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. *JMIR mHealth and uHealth*, 8(7):e17558.

Vasanti S. Malik, Walter C. Willett, and Frank B. Hu. 2013. Global obesity: trends, risk factors and policy implications. *Nature Reviews Endocrinology*, 9:13–27.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2024. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56:1–40.

Vinaytosh Mishra, Fahmida Jafri, Nafeesa Abdul Kareem, Raseena Aboobacker, and Fatma Noora. 2024. Evaluation of accuracy and potential harm of chatgpt in medical nutrition therapy-a case-based approach. *F1000Research*, 13:137.

Marc Mitchell and Lena Kan. 2019. Digital technology and the future of health systems. *Health Systems & Reform*, 5:113–120.

Andre Matthias Müller, Stephanie Alley, Stephanie Schoeppe, and Corneel Vandelanotte. 2016. The effectiveness of e-& mhealth interventions to promote physical activity and healthy diets in developing countries: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 13:109.

Farah Naja, Mandy Taktouk, Dana Matbouli, Sharfa Khaleel, Ayah Maher, Berna Uzun, Maryam Alameddine, and Lara Nasreddine. 2024. Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome. *European Journal of Clinical Nutrition*.

Valentina Natalucci, Federica Marmondi, Michele Biraghi, and Matteo Bonato. 2023. The effectiveness of wearable devices in non-communicable diseases to manage physical activity and nutrition: Where we are? *Nutrients*, 15:913.

Paweł Niszczota and Iga Rybicka. 2023. The credibility of dietary advice formulated by chatgpt: Robodiets for people with food allergies. *Nutrition*, 112:112076.

Mahmud Omar, Girish N Nadkarni, Eyal Klang, and Benjamin S Glicksberg. 2024. Large language models in medicine: A review of current clinical trials across healthcare applications. *PLOS Digital Health*, 3(11):e0000662.

Oladapo Oyebode, Chinenye Ndulue, Mona Alhasani, and Rita Orji. 2020. *Persuasive Mobile Apps for Health and Wellness: A Comparative Systematic Review*, volume 12064 LNCS, pages 163–181. Springer.

Vijaya Parameswaran, Jenna Bernard, Alec Bernard, Neil Deo, David Bates, Kalle Lyytinen, and Rajesh Dash. 2024. Optimizing large language models for interpreting american heart association dietary guidelines in nutrition education for cardiovascular disease prevention: A retrieval-augmented generation framework. *Circulation*, 150(Suppl_1):A4144884–A4144884.

Jung Ha Park, Ji Hyun Moon, Hyeon Ju Kim, Mi Hee Kong, and Yun Hwan Oh. 2020. Sedentary lifestyle: Overview of updated evidence of potential health risks. *Korean Journal of Family Medicine*, 41:365–373.

Yoobin Park, Brian P Don, Ashley E Mason, Aric A Prather, and Elissa S Epel. 2024. Daily social resources as a buffer against stress eating and its consequences. *Health Psychology*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Annisa Ristya Rahmanti, Hsuan Chia Yang, Bagas Suryo Bintoro, Aldilas Achmad Nursetyo, Muhammad Solihuddin Muhtar, Shabbir Syed-Abdul, and Yu Chuan Jack Li. 2022. Slimme, a chatbot with artificial empathy for personal weight management: System design and finding. *Frontiers in Nutrition*, 9.

William R. Rowley, Clement Bezold, Yasemin Arikan, Erin Byrne, and Shannon Krohe. 2017. Diabetes 2030: Insights from yesterday, today, and future trends. *Population Health Management*, 20:6–12.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Maksut Senbekov, Timur Saliev, Zhanar Bukeyeva, Aigul Almabayeva, Marina Zhanaliyeva, Nazym Aitenova, Yerzhan Toishibekov, and Ildar Fakhradiyev. 2020. The recent progress and applications of digital technologies in healthcare: A review. *International Journal of Telemedicine and Applications*, 2020:1–18.

Natalie Stein and Kevin Brooks. 2017. A fully automated conversational artificial intelligence for weight loss: Longitudinal observational study among overweight and obese adults. *JMIR Diabetes*, 2:e28.

Taylor N. Stephens, Angela Joerin, Michiel Rauws, and Lloyd N. Werk. 2019. Feasibility of pediatric obesity and prediabetes treatment support through tess, the ai behavioral coaching chatbot. *Translational Behavioral Medicine*, 9:440–447.

Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. 2004. Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544.

Michael Strange. 2024. Three different types of ai hype in healthcare. *AI and Ethics*, 4(3):833–840.

Laura P. Svetkey, Bryan C. Batch, Pao-Hwa Lin, Stephen S. Intille, Leonor Corsino, Crystal C. Tyson, Hayden B. Bosworth, Steven C. Grambow, Corrine Voils, Catherine Loria, John A. Gallis, Jenifer Schwager, and Gary B. Bennett. 2015. Cell phone intervention for you (city): A randomized, controlled trial of behavioral weight loss intervention for young adults using mobile technology. *Obesity*, 23:2133–2141.

Fawad Taj, Michel C A Klein, and Aart van Halteren. 2019. Digital health behavior change technology: Bibliometric and scoping review of two decades of research. *JMIR mHealth and uHealth*, 7:e13311.

Edmund R Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology*, 38(2):227–242.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty.

Annette Totten, Dana M. Womack, Marian S. McDonagh, Cynthia Davis-O'Reilly, Jessica C. Griffin, Ian Blazina, Sara Grusing, and Nancy Elder. 2022. Improving rural health through telehealth-guided provider-to-provider communication.

Vasileios Tsiantis, Dimitrios Konstantinidis, and Kosmas Dimitropoulos. 2024. Chatgpt in nutrition: Trends challenges and future directions. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 548–553. ACM.

Floor van Meer, Lisette Charbonnier, and Paul AM Smeets. 2016. Food decision-making: effects of weight status and age. *Current diabetes reports*, 16:1–8.

Maria F Vasiloglou, Jane Fletcher, and Kalliopi-Anna Poulia. 2019. Challenges and perspectives in nutritional counselling and nursing: a narrative review. *Journal of clinical medicine*, 8(9):1489.

Laura Vearrier, Kyle Rosenberger, and Valerie Weber. 2018. Use of personal devices in healthcare: Guidelines from a roundtable discussion. *Journal of Mobile Technology in Medicine*, 7:27–34.

Laura M. Vowels, Rachel R.R. Francois-Walcott, and Joëlle Darwiche. 2024. Ai in relationship counselling: Evaluating chatgpt's therapeutic capabilities in providing relationship advice. *Computers in Human Behavior: Artificial Humans*, 2:100078.

Aleksandra Vrkatić, Maja Grujičić, Jelena Jovičić-Bata, and Budimka Novaković. 2022. Nutritional knowledge, confidence, attitudes towards nutritional care and nutrition counselling practice among general practitioners. In *Healthcare*, volume 10, page 2222. MDPI.

Brian Wansink and Jeffery Sobal. 2007. Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior*, 39:106–123.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *International Conference on Learning Representations (ICLR)*.

B. D. Weiss. 2005. Quick assessment of literacy in primary care: The newest vital sign. *The Annals of Family Medicine*, 3:514–522.

Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, and Amir M Rahmani. 2024. Chatdiet: Empowering personalized nutrition-oriented food recommender chatbots through an llm-augmented framework. *Smart Health*, 32:100465.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint. *Journal of Medical Internet Research*, 22:e22845.

A Trial Checks

Throughout the duration of the trial, the chatbot performed several daily checks on participants to ensure and encourage task adherence:

- At 10am, the chatbot checked for any **abnormalities** in the previous day's food logs, considering a combination of objective measures and heuristic decisions. The abnormalities found could be to do with mistakes in logging, or diary incompleteness. If it noticed an abnormality, the chatbot sent a message asking the user to double check their food diary on MyFitnessPal. These checks investigated whether the user's food diary from the previous day:
  - was empty;
  - consisted of less than half their calorie goal;
  - consisted of more than twice their calorie goal;
  - had less than four food items recorded;
  - contained a particular food item with a recorded amount of more than one kilogram, two litres, or six cups.
- At 4pm, the chatbot sent a reminder to any user with an **empty diary** that day, i.e. anyone who had yet to log a food item.
- At 6pm, if the user had **not interacted** with the chatbot in the last **36 hours**, the chatbot encouraged them to initiate a conversation.

B Technical Details

For fine-tuning the nutritional counselling models, we had access to NVIDIA A40 GPUs with 48GB of VRAM. We applied 4-bit quantisation to optimise memory usage and enable efficient training of the larger of the chosen models (Gemma 7B, Llama 3 8B, and Mistral 7B). The prompts are shown in Table B.1 and the training configurations are outlined in Table B.2. We monitored validation loss during training and selected the best-performing checkpoints. Batch sizes and training durations were adjusted to match the capabilities of the available GPUs.

Towards LLM-based rephrasing, we selected the models according to our technical constraints, as with the nutritional counselling models, but with extra consideration of the inference speeds. This was initially a major limitation despite a simpler prompting approach on the machine deploying the chatbot. Running full models on the available Quadro RTX 5000 GPU resulted in average decoding times of over 30 seconds per message—too slow for a real-time chatbot. We explored several optimisation strategies, including 4-bit quantisation, KV-caching, `torch.compile`, flash attention. While most techniques yielded minimal improvements, combining the 4-bit quantised Llama 3 with the Ollama API drastically reduced decoding time to around 1.1 seconds per message, making real-time deployment viable. Results from an initial prompt (Figure 3) on 19 example messages with varying intents and complexity are outlined in Table B.3. Among the three models, Mistral was quickly ruled out due to inconsistent outputs and occasional language mixing. Gemma performed better, but tended to omit key information and exhibited slower inference times. Llama 3 8B emerged as the most consistent, producing diverse and accurate rephrasings, and was selected as the production model.

C Rephrasing Evaluation

In order to verify the suitability of the rephrasing model, we ran an intrinsic evaluation with human crowd workers to compare the templated responses against the rephrased ones. In this experiment, our objective was to evaluate the preference of chatbot users towards the original templated responses of the baseline chatbot or the responses that have been rephrased by the prompted model. Furthermore, we aimed to determine which response sounded more natural to users and whether users could distinguish any difference in meaning between the two texts. Only a few semantic mismatches were reported, typically due to numerical misinterpretation in isolated cases.

The evaluation took the form of an annotation task that presented a random sample of 12 pairs of templated responses and rephrased outputs, with accompanying conversational context, covering a diverse range of user queries and scenarios. These text pairs are included in our repository. The participants were shown these text pairs as a sequence of twelve questions, including an attention question designed to make sure that they were completing the task according to the instructions. To each presented text pair, they had to answer three subquestions:

1. Which response do you prefer?
2. Which response sounds more natural?
3. Do both responses have the same meaning?

Text	Prompts
REFL	You are an expert dietitian. Below is a struggle your client is experiencing. Summarize what the problem is about or infer what they mean. Do not assume their feelings. ### Struggle: `[STRUGGLE]` ### Reflection:
COMF	You are an expert dietitian. Below is a struggle your client is experiencing. Tell them that the situation is not unrecoverable, normalize the situation or make them feel understood. Do not normalize dangerous behaviours in a way that explicitly encourages your client to commit them. ### Struggle: `[STRUGGLE]` ### Comfort:
REFR	You are an expert dietitian. Below is a struggle your client is experiencing. Show a benefit to the struggle that they did not consider or find something about the struggle to be grateful for. ### Struggle: `[STRUGGLE]` ### Reframing:
SUGG	You are an expert dietitian. Below is a struggle your client is experiencing. Tell the person how to change their habit to improve or suggest an alternative helpful activity. ### Struggle: `[STRUGGLE]` ### Suggestion:

Table B.1: The instruction prompts used to fine-tune the instruction-tuned models.

Model	Batch	Epochs	LR	Optimiser
`GPT-2 small`	8	10	5e-5	`AdamW`
`GPT-2 medium`	4	10	5e-5	`AdamW`
`BabyLlama`	8	20	5e-5	`AdamW`
`FLAN-T5 base`	8	30	1e-4	`AdamW`
`Gemma 2B`	2	3	2e-4	`paged_adamw_8bit`
`Gemma 7B`	2	3	2e-4	`paged_adamw_8bit`
`Mistral 7B`	2	3	2e-4	`paged_adamw_8bit`
`Llama 3 8B`	2	3	2e-4	`paged_adamw_8bit`

Table B.2: Parameters for the fine-tuning of the nutritional counselling models. LR = learning rate.

Technique	Llama 3	Gemma	Mistral
Full model	29	48	31
4-bit model	6.3	7.6	11
Unsloth	34	—	—
Ollama	1.1	1.1	—

Table B.3: Average decoding time in seconds per example for each of the 19 example messages rephrased by the models using various methods.

If the participant responded with "No", they could optionally provide a reason.

Participants of this annotation task were sourced on Prolific (https://www.prolific.com), under conditions that they were primarily an English speaker.

Text pair	Preferred	More natural
1	70%	85%
2	75%	85%
3	65%	65%
4	55%	65%
5	60%	55%
6	75%	90%
7	55%	60%
8	65%	75%
9	60%	75%
10	85%	80%
11	55%	65%

Table C.1: Results of the human evaluation of the rephrased responses in comparison to the templated responses in terms of proportion of participants in favour of the rephrased response.

In total, 23 crowd workers on the platform completed the task, of which 20 passed the attention question to be considered in our analysis.

The results of the task are displayed in Tables C.1 and C.2. Overall, the findings suggest that while participants generally preferred and found the rephrased responses more natural compared to the templated responses, the preference was not overwhelmingly strong, with some text pairs showing a narrower margin of preference. While the templated outputs featured more structured formatting through the use of newlines and bolding, the rephrased outputs leaned toward a more conversational style, often incorporating emojis as instructed. This difference in presentation may have influenced participant preferences and contributed to the higher perceived naturalness of the rephrased responses.

D Ethics Details

Ethical approval was obtained from the University of Aberdeen as well as ethical advisors at Charles University, and informed consent was obtained twice from all trial participants during onboarding: once during registration and again at onboarding, ensuring participants fully understood the study tasks (after group assignment) and data usage. Participants had the right to withdraw at any point before data analysis began, and any data from those who withdrew early was excluded from analysis. As compensation, participants were given online gift vouchers (Alza or Amazon) worth €8, AU$13, or 200CZK for each completed week of the first six weeks of the experiment, and €16, AU$26, or 400CZK for the seventh (and final) week of the experiment, if completed.

E Technical Challenges During the Trial

During the initial setup of the model for the user trial, the model was trained on the first safe supportive text, prioritising the candidates provided by experts. However, part-way through the user trial, we identified an oversight: we had inadvertently ignored many other safe candidates in training that make up about seven times the number of training examples in the initial fine-tuning. That is, in the original model, a struggle was trained with a single corresponding text category, but, in fact, there are up to ten generated candidates and up to three expert-provided texts that can be used for training. To address this, the model was immediately retrained on the full set of safe candidates and swapped out with the original model serving the chatbot in the user trial in time for the fifth week since the beginning of the trial.

This approach provided the opportunity to compare the performance between the original and updated models. At the conclusion of the trial, we asked users if they observed any difference in the performance of the nutritional counselling model since the trial start. About 28% of participants (n=8) agreed that they noticed an improvement, while 48% (n=13) provided a neutral response, indicating no significant change in their experience. The remaining 24% (n=7) disagreed. These results imply that the additional training data may not have led to substantial improvements in model performance. However, it is important to consider that the declining use of the "advice" feature over time, like the general decrease in chatbot interactions, may have influenced these perceptions.

The trial also faced several other technical challenges. The most significant was a temporary disruption in access to MyFitnessPal data between Weeks 4 and 5 (June 18–21), which prevented the chatbot from delivering personalised dietary insights. During this time, the nutritional counselling feature remained active. Users were informed of the issue and encouraged to continue logging meals independently. Full functionality was restored by June 21, and daily interaction requirements were relaxed to accommodate the interruption. Other disruptions that occurred were infrequent and resolved promptly.

	Texts	Explanation
1	1	The "on average" section had the numbers reversed.
2	1	One has a 40% deficit and one has a 60% one
	9	One asks to be let know what changed one does not
3	4	[The templated response] states 'to see how different foods contribute to your intake' but [the rephrased response] states 'foods that contributed to your daily intake'
	6	[The templated response] is referring to safety but [the rephrased response] is referring to the accuracy of the results
4	9	—
5	1	They differed in the average protein that the user consumed, [the rephrased response] claimed they only hit 40% of their protein goal while [the templated response] claimed that they hit 60%.
6	1	Numbers juxtapositioned

Table C.2: Results of the human evaluation of the rephrased responses in comparison to the templated responses in terms of the reporting of differences in meaning.

F Trial Demographics

Demographic data collected at registration included age, gender identity, educational background, occupation, ethnicity, native country, English proficiency, and nutritional literacy captured through Pfizer's NVS questionnaire (Weiss, 2005). The collected statistics are illustrated across Figures F.1 and F.2.

The final study group was predominantly female, well-educated, ethnically diverse, and with adequate English skills and nutritional literacy for engaging with the chatbot and interpreting dietary insights.

G Trial Dietary Results by Weight Goal

To explore variation in diet outcomes according to the different weight goals, we analysed absolute distance and goal percentage trends among participants aiming to lose, maintain, or gain weight by fitting regression lines to each study group (Figures G.3 to G.8). Those seeking to lose weight and receiving nutritional counselling showed greater improvements in goal adherence—especially for energy and protein intake—than other groups. However, participants aiming to gain weight saw poorer adherence across most nutrients in the nutritional counselling group. These subgroup trends suggest that the nutritional counselling feature may have had uneven effects, depending on nutritional goals.

H Emotional Well-being Metrics

In an analysis of the result in **positive affect** by weight goal subgroups (Figure H.11), we observed a clear divergence in the effectiveness of interventions across groups. Participants with a goal to **lose weight** demonstrated the most consistent and notable improvements in positive affect scores, particularly in `REPHRASED`, where scores steadily increased from Week 1 to Week 7. In contrast, those aiming to **maintain weight** showed more stable and less pronounced changes across all groups. While `REPHRASED` still outperformed `BASELINE` and `FULL`, the magnitude of change for those aiming to maintain their weight was less dramatic compared to those aiming to lose some, suggesting that participants maintaining weight may experience fewer fluctuations in emotional or motivational states due to a less urgent goal. The subgroup of participants looking to **gain weight**, however, presented more varied outcomes. Positive affect scores fluctuated considerably across weeks, with `FULL` showing the steepest increases toward the end of the study.

An analysis of **positive affect** across the specific emotions (Figure H.12 revealed notable differences between them. While some positive emotions such as "Alert" and "Inspired" exhibited slight improvements, others, like "Determined" and "Enthusiastic," showed either stagnation or a gradual decline over the weeks. This variability suggests that the nutritional counselling intervention in this work may not have effectively addressed the emotional

dimensions critical for sustained engagement and motivation.

Again, we analysed the subgroups with different weight goals for **negative affect** (Figure H.15). Among participants who aimed to **lose weight**, those in FULL experienced minimal reductions in negative affect compared to those in BASELINE and REPHRASED, both of which showed significant declines. With participants who wanted to **maintain weight**, FULL again failed to show improvement, with a relatively flat trend line, while the other two groups showed consistent reductions. For participants aiming to **gain weight**, FULL exhibited the least improvement, with scores fluctuating without a clear downward trend.

Examining the individual **negative emotions** (Figure H.16) highlights further limitations of nutritional counselling for FULL. The emotions of "Afraid", "Scared", and "Upset" showed little to no improvement in FULL, while the other two groups experienced gradual reductions over the intervention period. Nervousness scores in FULL remained stable or even increased slightly, unlike REPHRASED, which showed a notable decline in this emotion by the end of the study. Despite some fluctuation, there was no significant improvement in distress scores for FULL, in stark contrast to REPHRASED, which showed a steep decline.

I User Engagement Metrics

To investigate the effect of the rephrasing of templated responses that was present in both REPHRASED and FULL, we looked at how user engagement differed between these groups and BASELINE. As engagement metrics, we used the interactions (i.e. individual messages from the user), conversations (i.e., a sequence of interactions with responses within five minutes of each other), and days that users interacted with the chatbot, considering the total number and distribution over the trial duration.

As shown in Figures I.1 and I.2, the mean number of interactions per day and week declined consistently over time for all groups, indicating a natural decrease in user engagement as the intervention progressed. However, FULL consistently demonstrated the highest number of interactions overall, likely reflecting the additional "advice" feature available exclusively to this group, as well as the prompt to use it every week. In contrast, the rephrased responses alone in REPHRASED did not appear to significantly affect engagement metrics over time compared to BASELINE.

The mean number of conversations per week (Figure I.3) presented more mixed results. While FULL maintained the highest number of conversations, as highlighted by the total number of conversations in Figure I.6, the differences were not as pronounced as the total number of interactions (Figure I.5). The lack of notable differences between BASELINE and REPHRASED suggests that the increase in overall conversations observed in FULL was also driven primarily by the additional "advice" feature, rather than the rephrased responses shared by REPHRASED and FULL. Furthermore, FULL did not engage with the chatbot on more days per week than the other groups (Figure I.4), so there was no clear evidence that the rephrased responses had any consistent impact on this metric.

These findings suggest that while the "advice" feature and the weekly prompt in FULL contributed to overall engagement, rephrased responses did not significantly influence the number of conversations or the frequency of chatbot use.

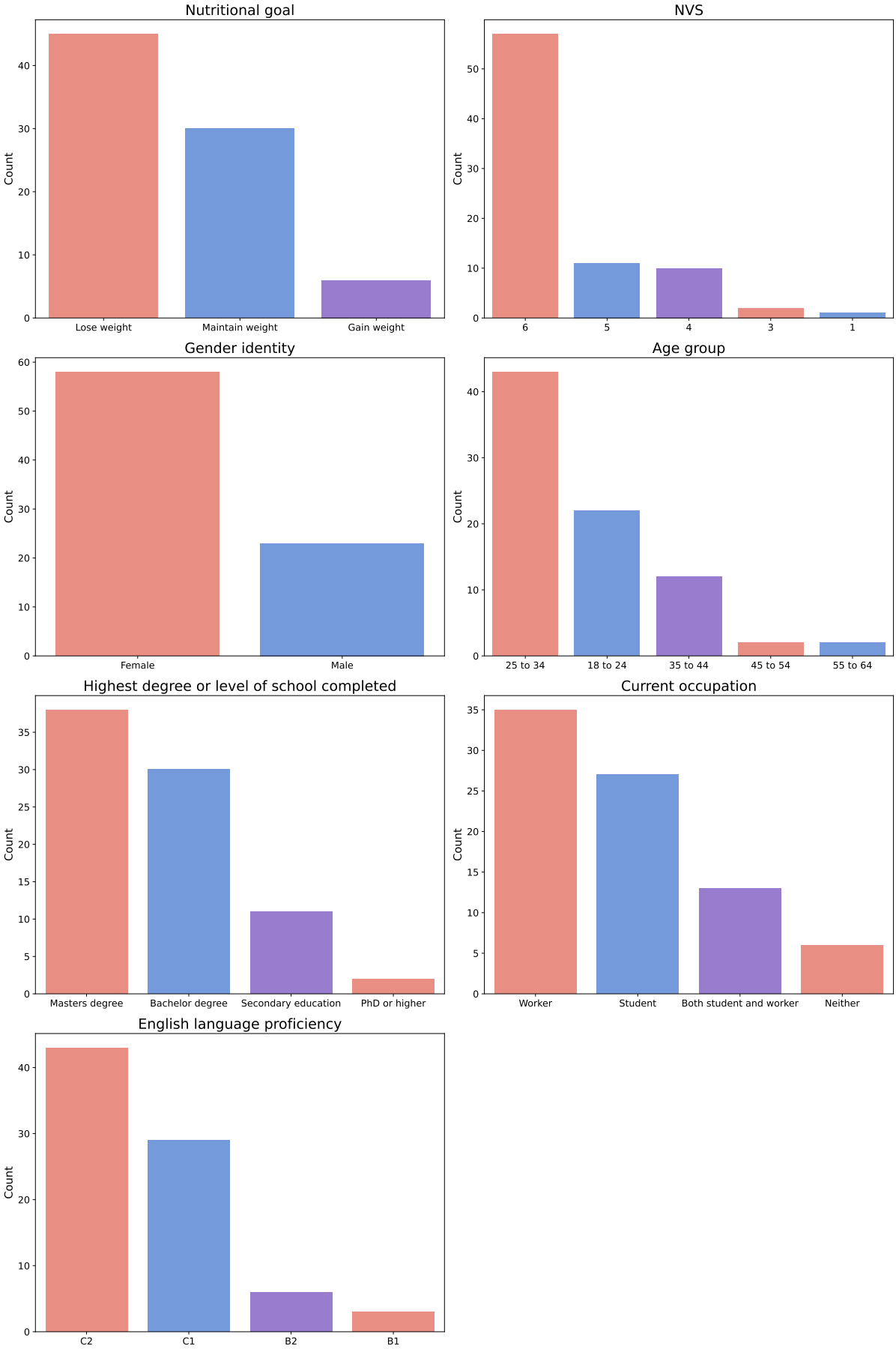


Figure F.1: Graphics showing collected demographic data of participants (part 1).

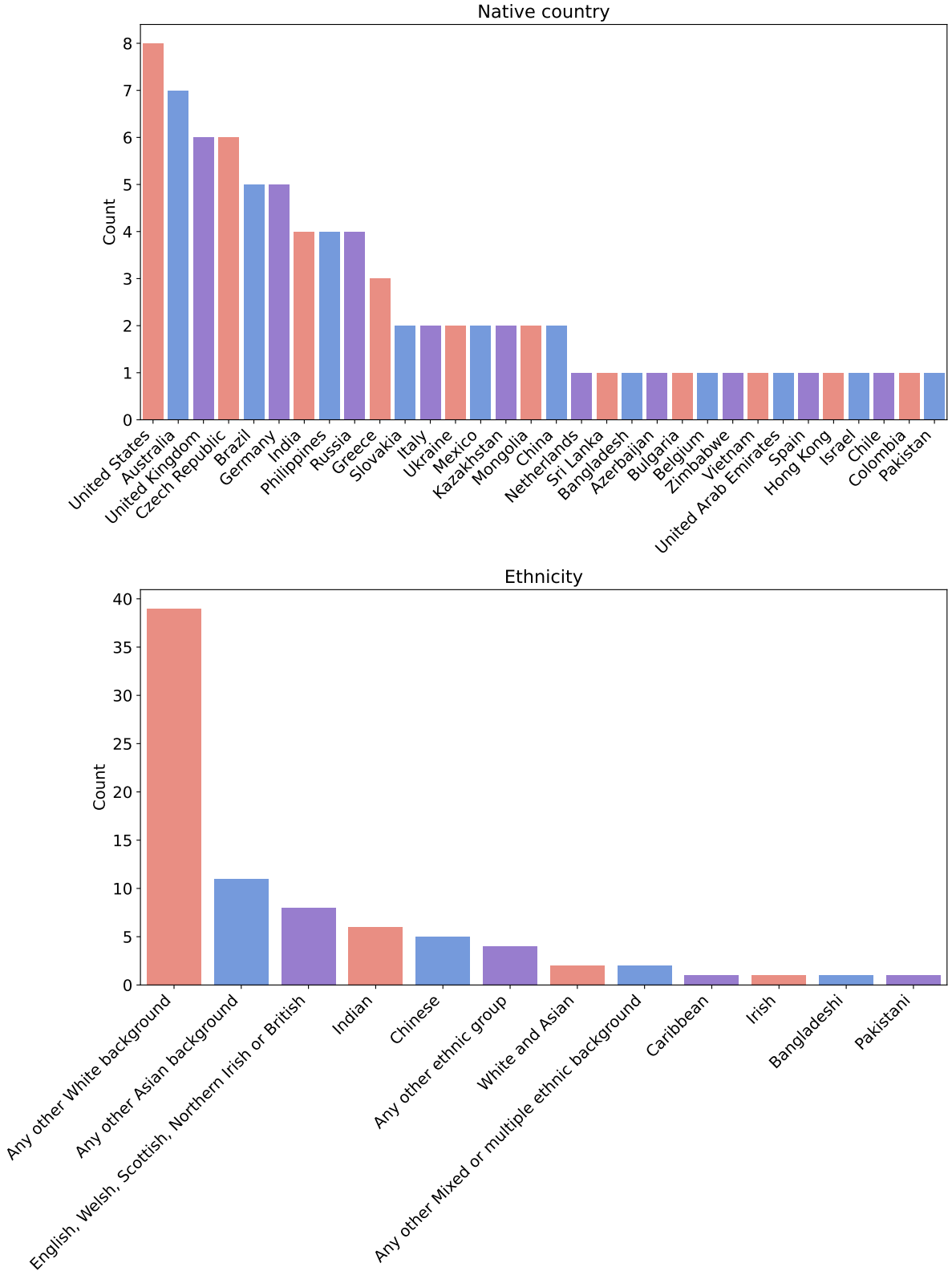


Figure F.2: Graphics showing collected demographic data of participants (part 2).

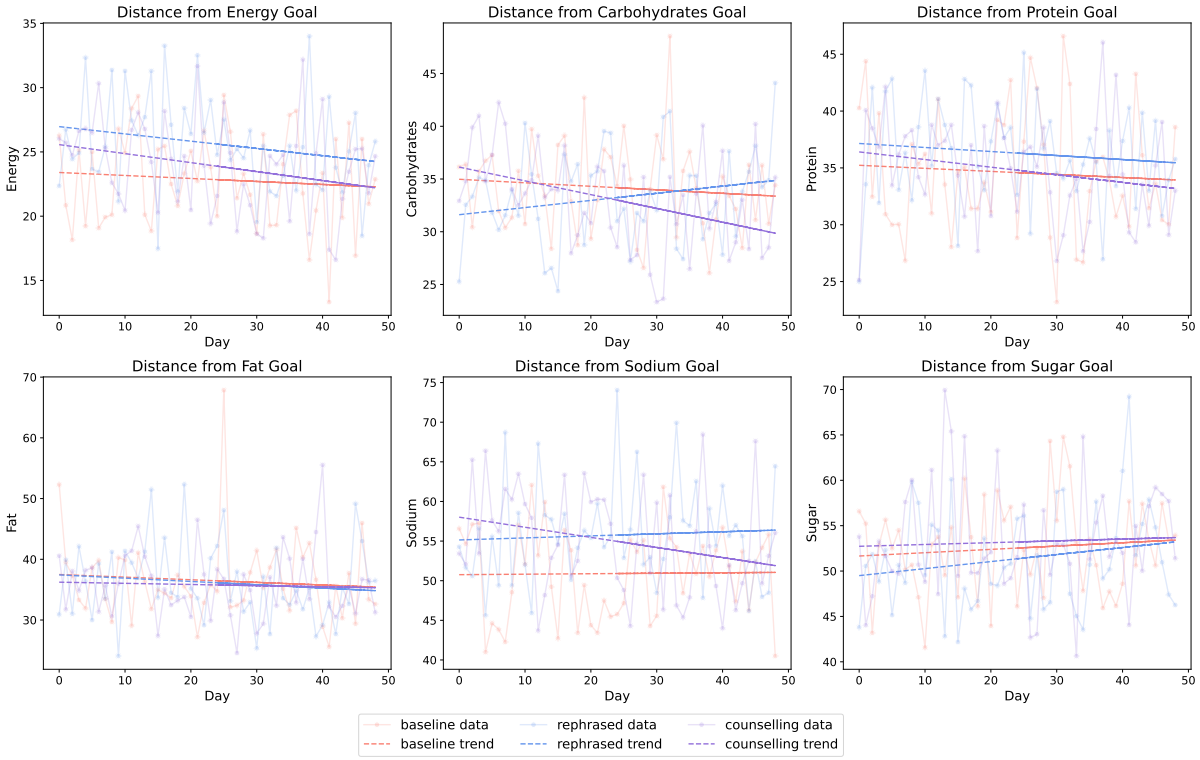


Figure G.1: Overall absolute percentage distance from MyFitnessPal intake goals.

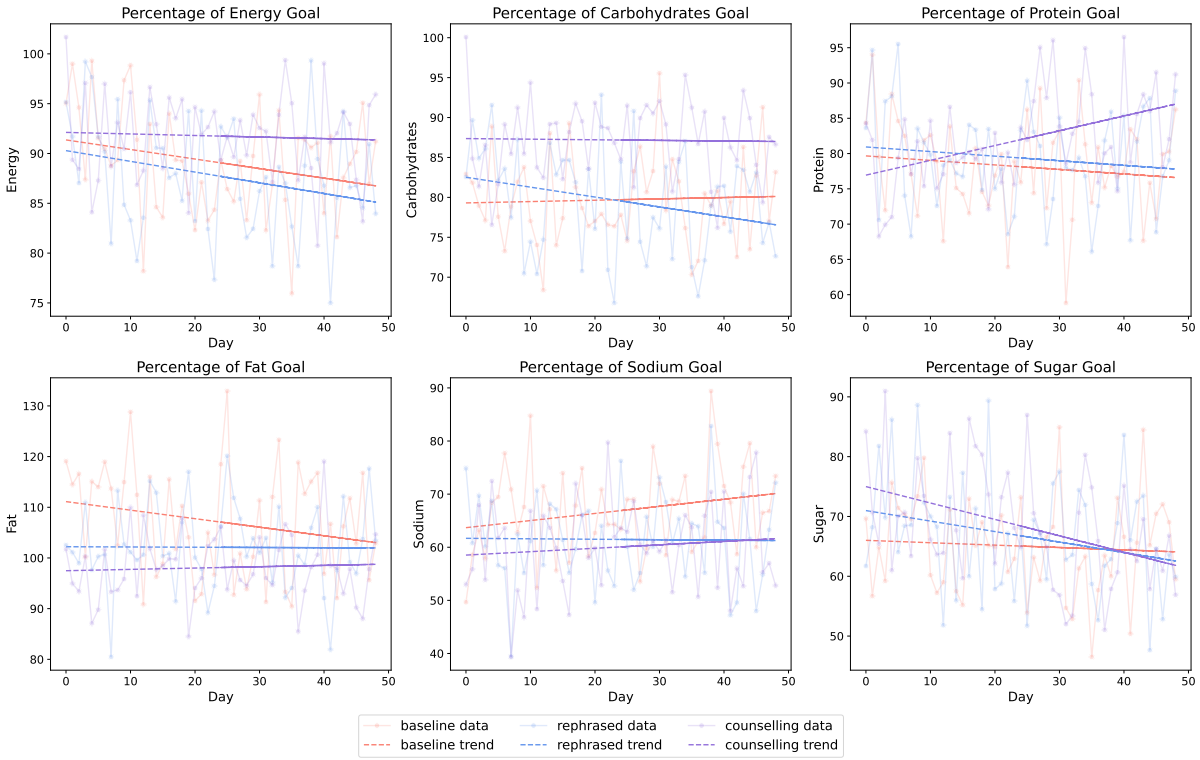


Figure G.2: Overall goal percentage of MyFitnessPal intake goals.

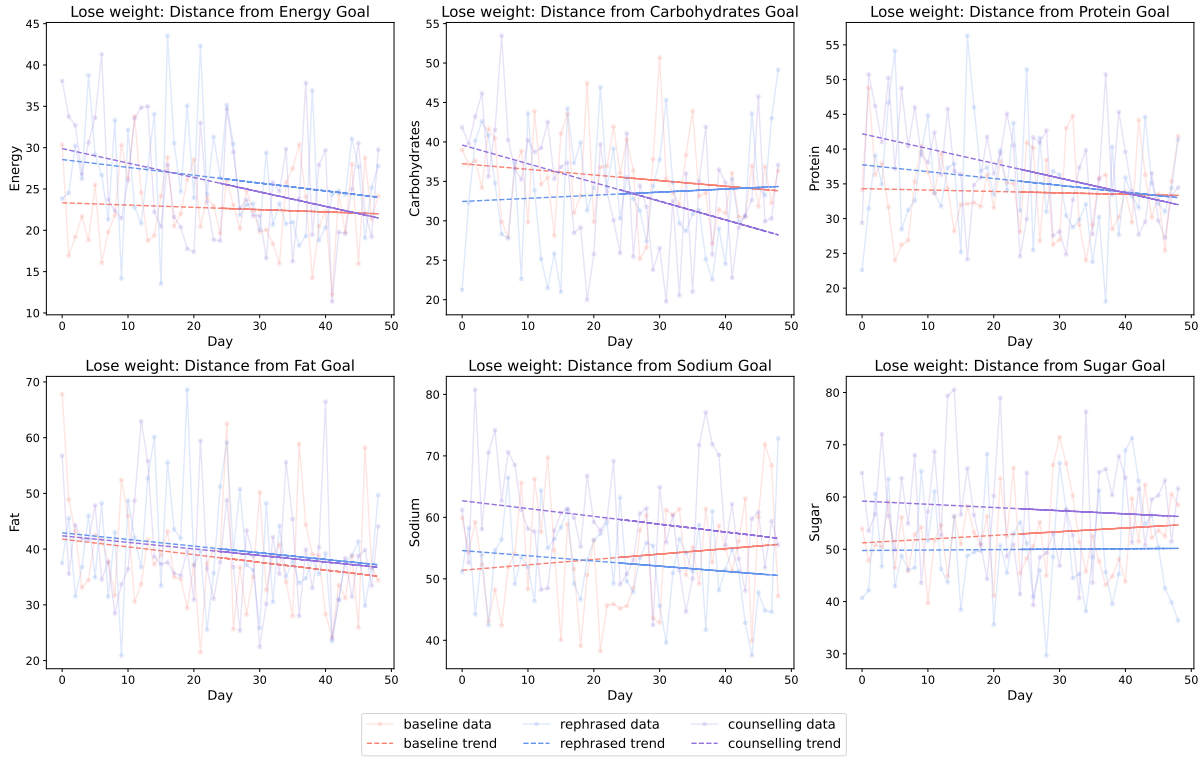


Figure G.3: Absolute percentage distance from MyFitnessPal intake goals for participants aiming to lose weight. LW

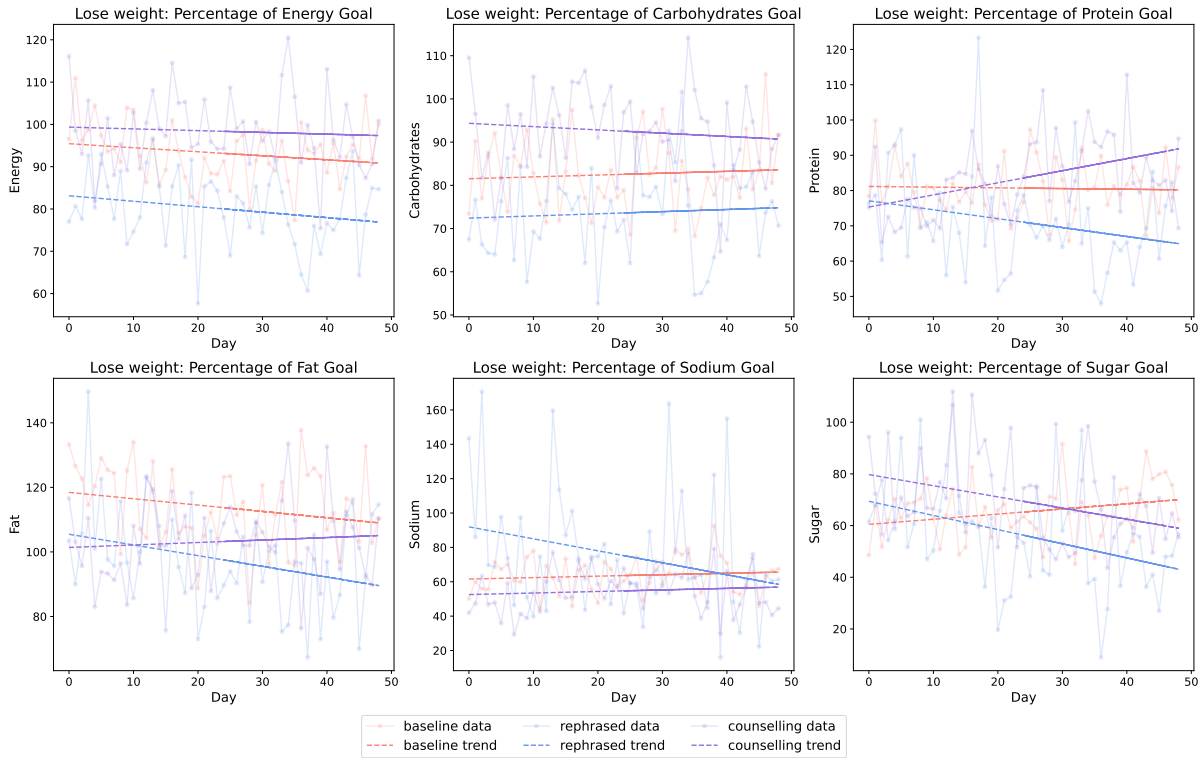


Figure G.4: Goal percentage of MyFitnessPal intake goals for participants aiming to lose weight.

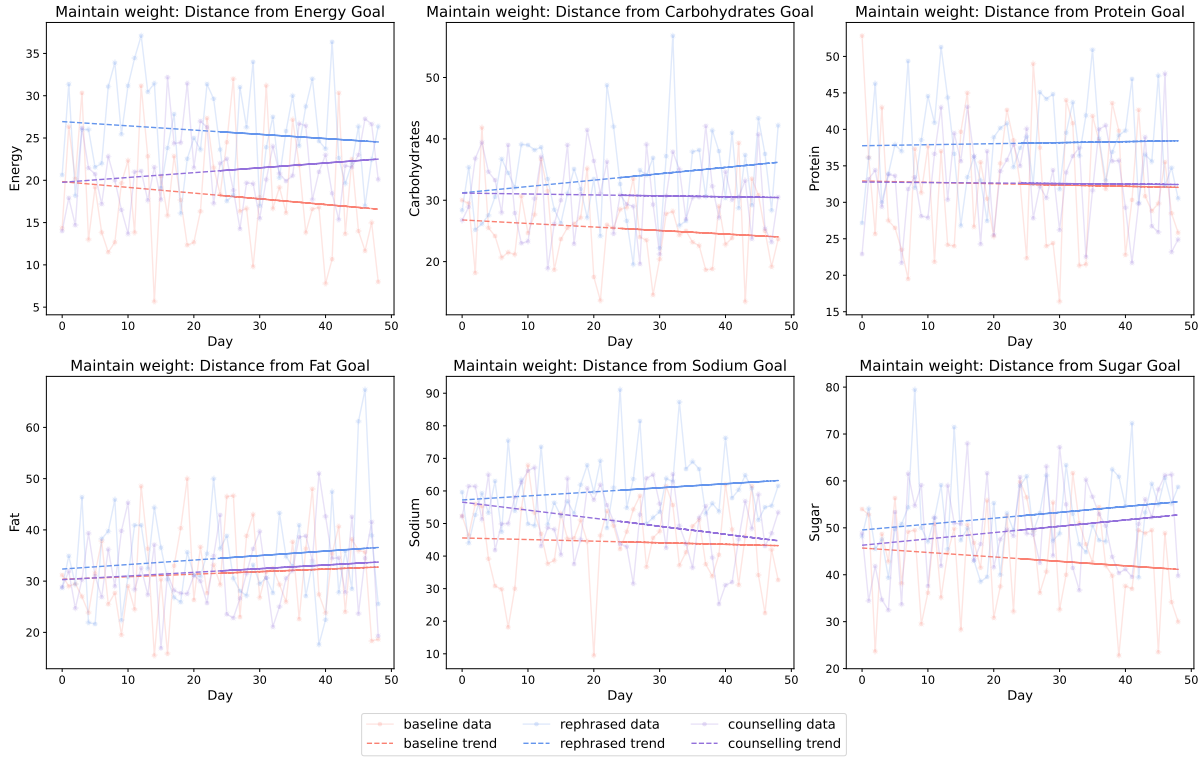


Figure G.5: Absolute percentage distance from MyFitnessPal intake goals for participants aiming to maintain their weight.

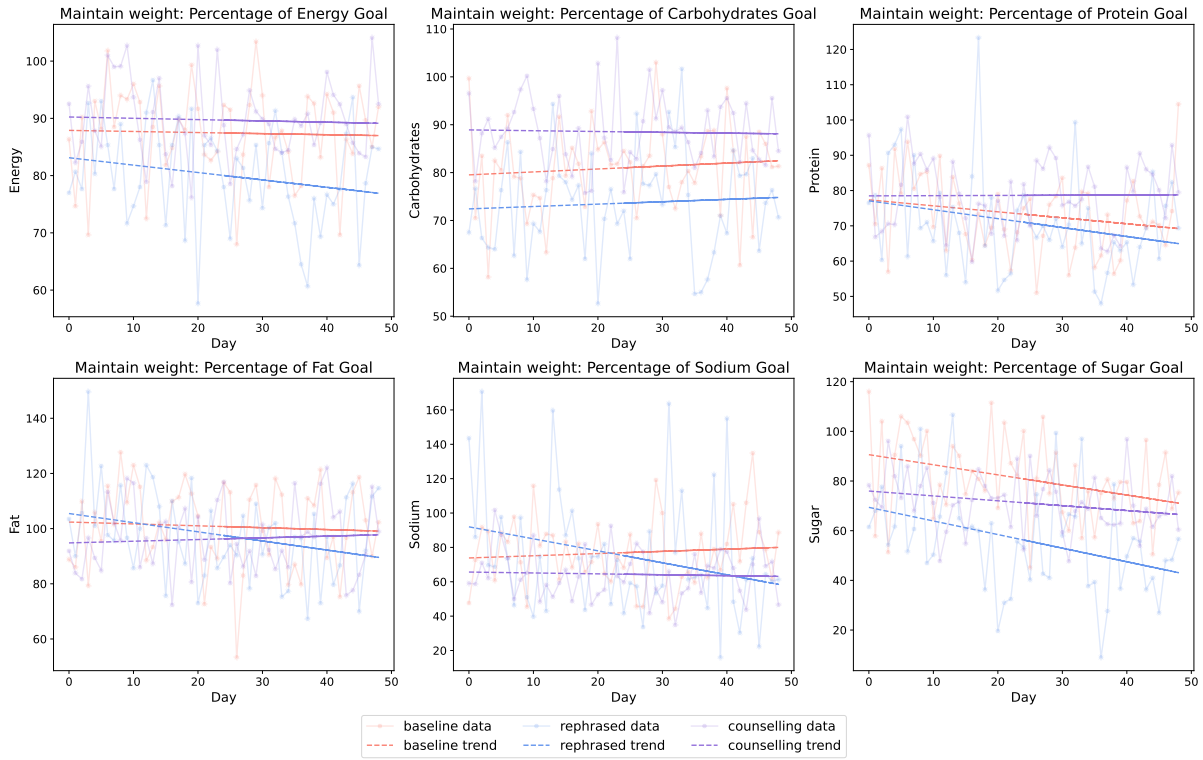


Figure G.6: Goal percentage of MyFitnessPal intake goals for participants aiming to maintain their weight.

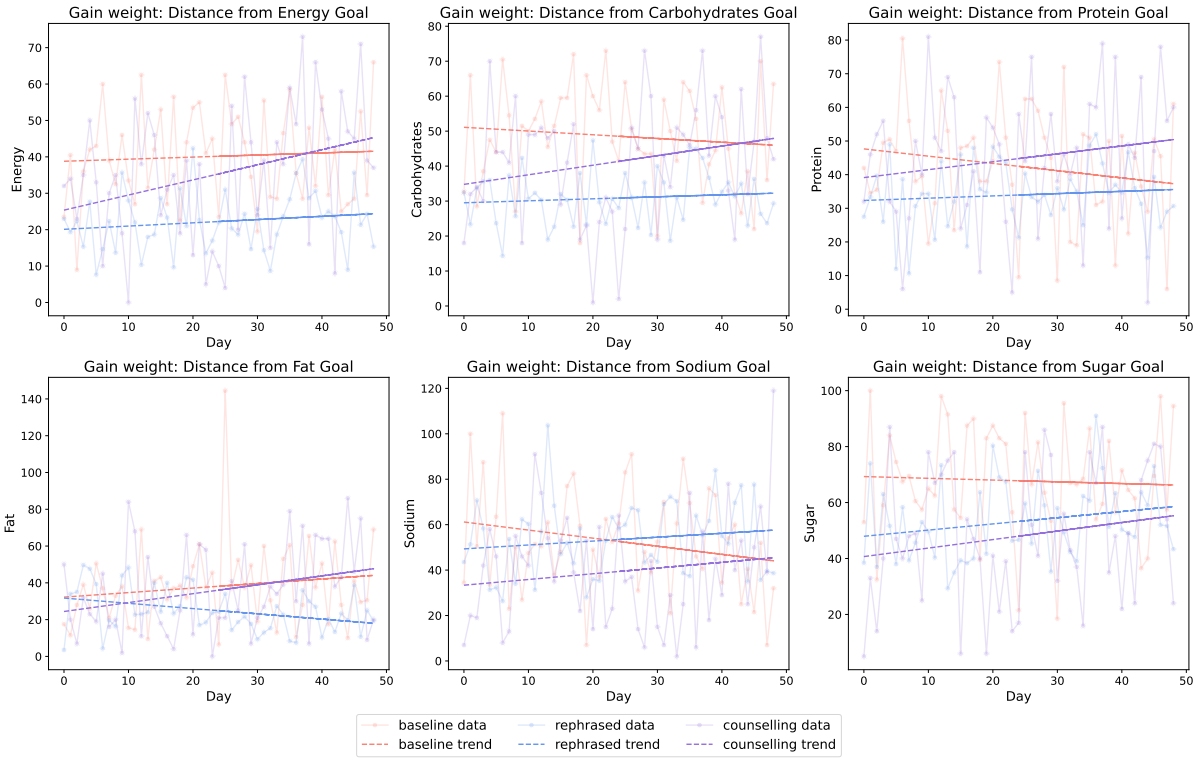


Figure G.7: Absolute percentage distance from MyFitnessPal intake goals for participants aiming to gain weight.

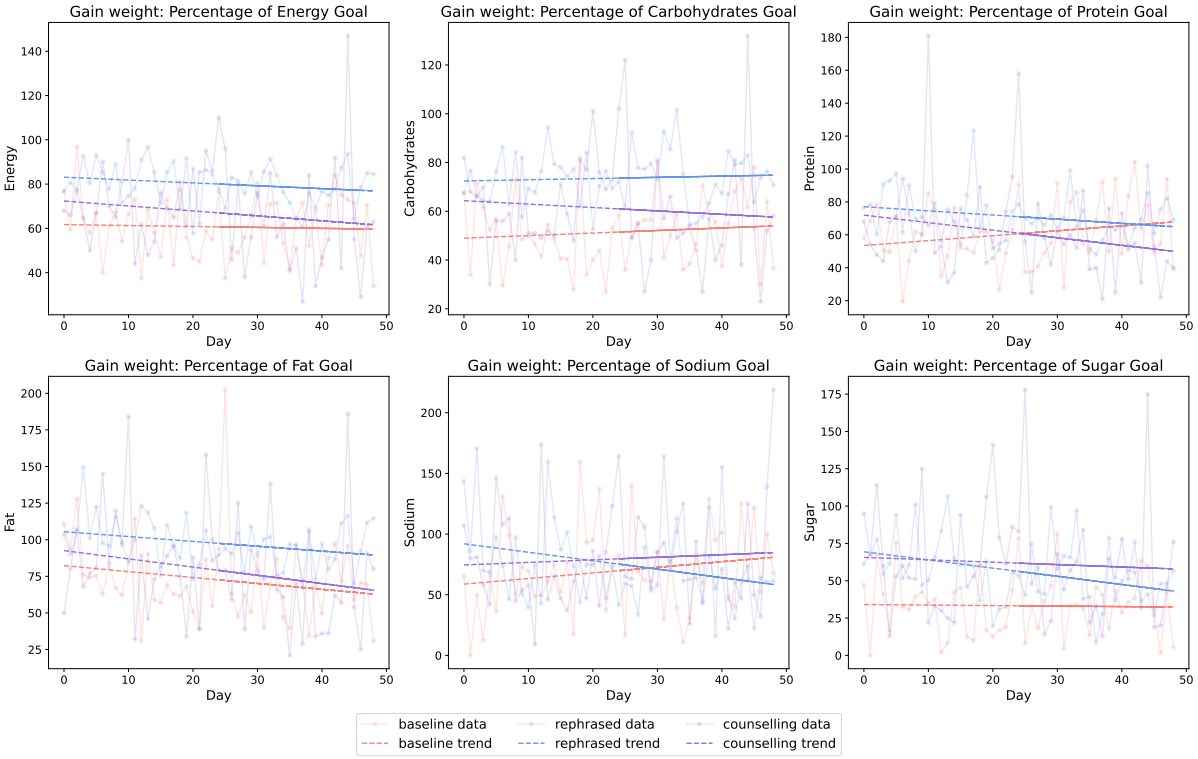


Figure G.8: Goal percentage of MyFitnessPal intake goals for participants aiming to gain weight.

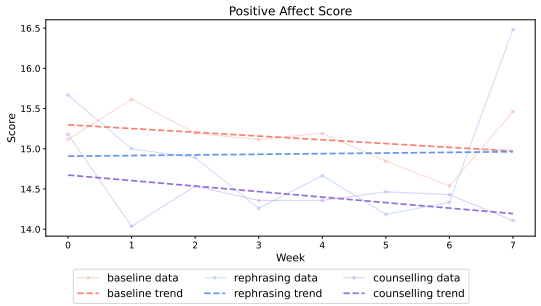


Figure H.9: Positive affect score.

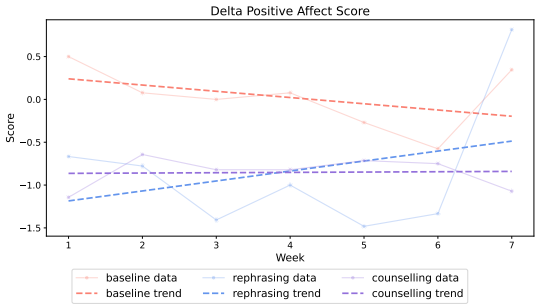


Figure H.10: Delta positive affect score.

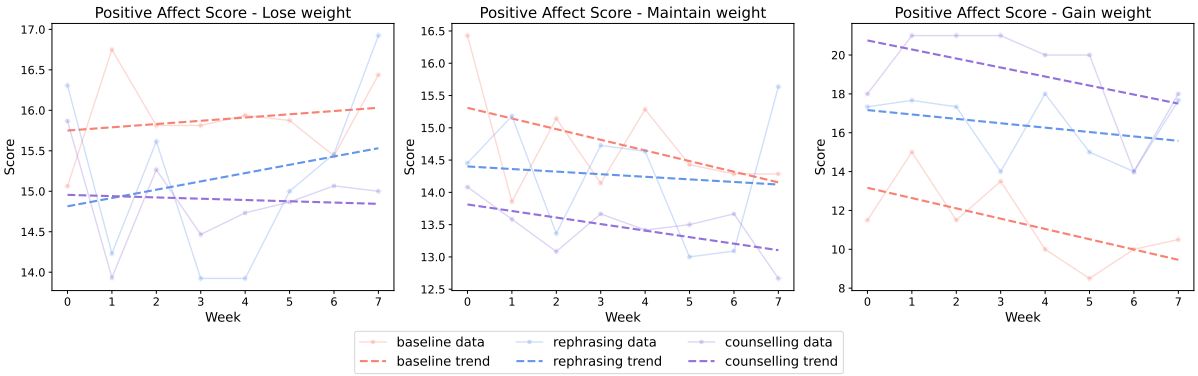


Figure H.11: Positive affect score by weight goal.

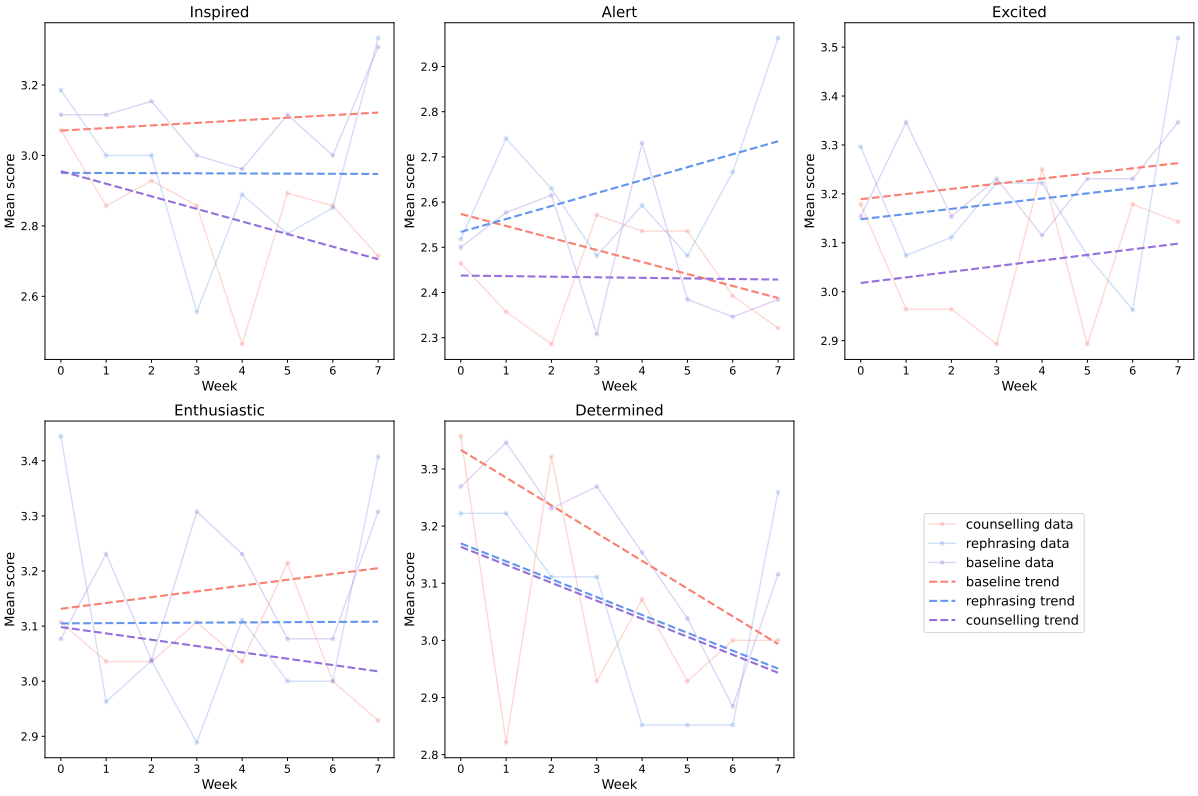


Figure H.12: Positive affect score by emotion.

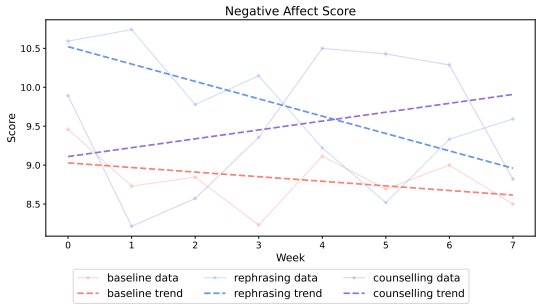


Figure H.13: Negative affect score.

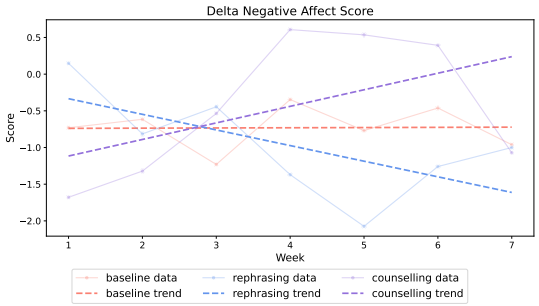


Figure H.14: Delta negative affect score.

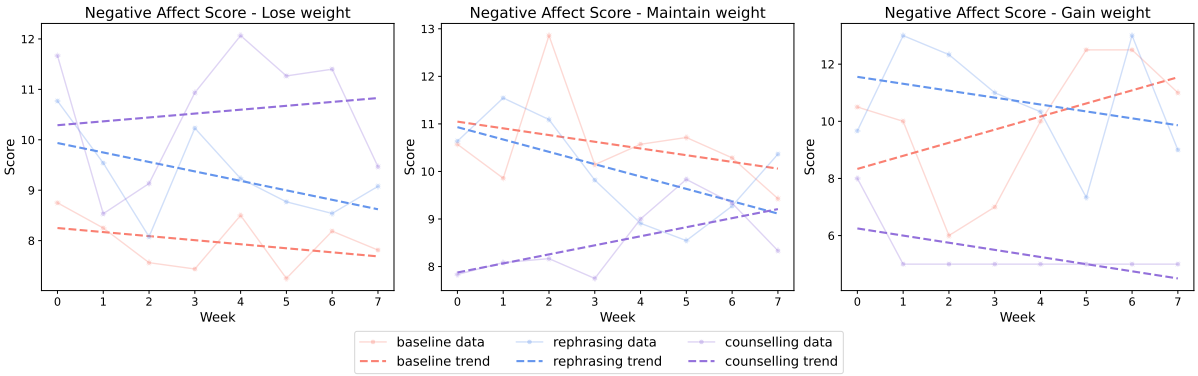


Figure H.15: Negative affect score by weight goal.

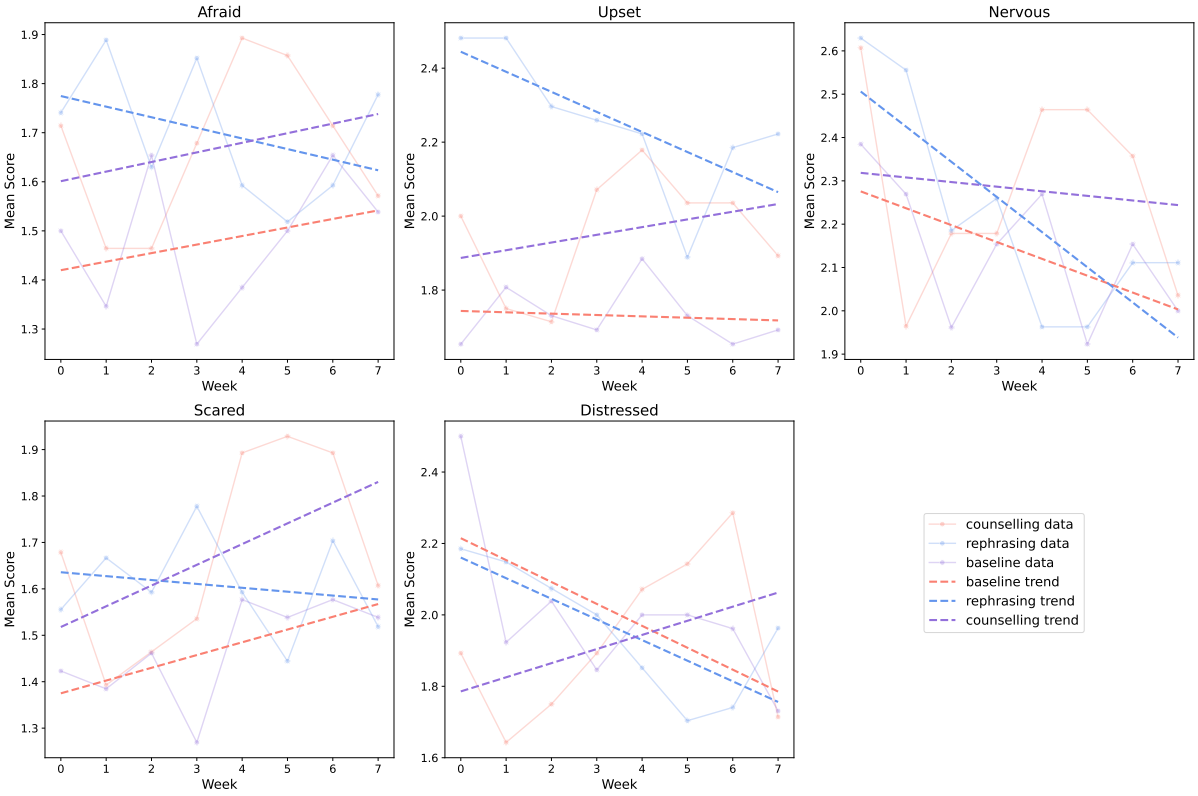


Figure H.16: Negative affect score by emotions.

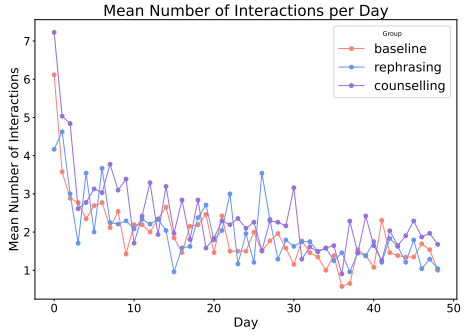


Figure I.1: Mean number of interactions per day.

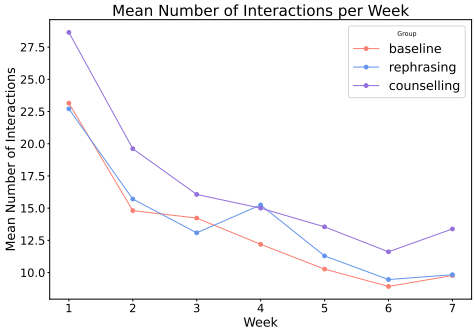


Figure I.2: Mean number of interactions per week.

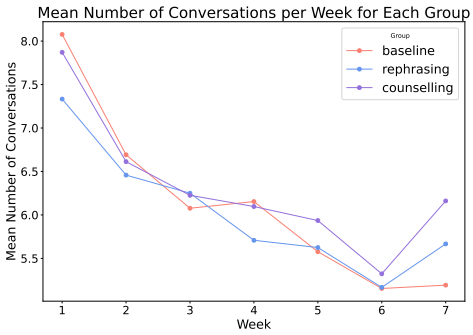


Figure I.3: Mean number of conversations per week.

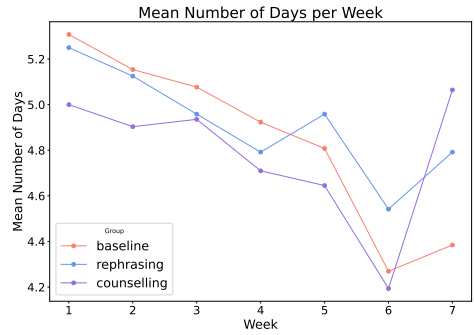


Figure I.4: Mean number of days users interacted with the chatbot per week.

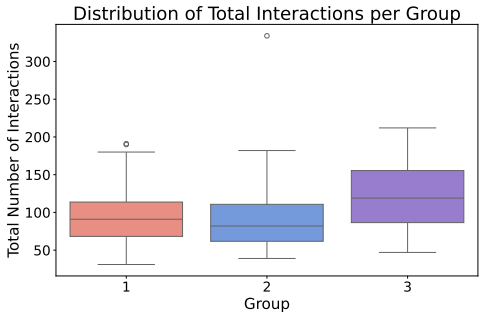


Figure I.5: Total number of interactions per group.

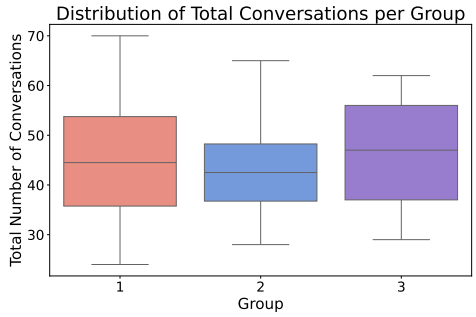


Figure I.6: Total number of conversations per group.