Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models

Georg Ahnert¹, Max Pellert², David Garcia^{2,3}, Markus Strohmaier^{1,4,3}

¹University of Mannheim, ²University of Konstanz, ³CSH Vienna, ⁴GESIS - Leibniz Institute for the Social Sciences

Abstract

This paper proposes temporally aligned Large Language Models (LLMs) as a tool for longitudinal analysis of social media data. We fine-tune Temporal Adapters for Llama 3 8B on full timelines from a panel of British Twitter users, and extract longitudinal aggregates of emotions and attitudes with established questionnaires. We validate our estimates against representative British survey data and find strong positive, significant correlations for several collective emotions. The obtained estimates are robust across multiple training seeds and prompt formulations, and in line with collective emotions extracted using a traditional classification model trained on labeled data. To the best of our knowledge, this is the first work to extend the analysis of affect in LLMs to a longitudinal setting through Temporal Adapters. Our work enables new approaches towards the longitudinal analysis of social media data.

Introduction

Problem A number of recent studies has used Large Language Models (LLMs) to perform *cross-sectional* surveys *in silico* (e.g., Argyle et al. 2023; Bisbee et al. 2023; Durmus et al. 2023), modeling human-based survey responses by utilizing LLMs. While showing great promise, such prompt-based in silico surveys have not been aligned well from a temporal perspective, meaning that LLM training data and survey fieldwork periods typically did not meaningfully overlap in terms of time. Independently, many studies have investigated affect aggregates from social media data, in particular collective emotions (e.g., Golder and Macy 2011; Pellert et al. 2022; Metzler et al. 2023). However, these methods typically rely on expensively curated dictionaries or on labeled data.

Approach In this paper, we present a novel method for extracting longitudinal affect aggregates by using *Temporal Adapters* fine-tuned on user-generated data from social media, as illustrated in Figure 1. We base our analysis on a panel of 24,000 British Twitter users and obtain their full timelines from November 2019 to June 2020. First, we fine-tune Temporal Adapters for Llama 3 8B on 7-day subsets of the Twitter data. Second, we prompt the fine-tuned model with multiple established questionnaires (YouGov 2024a,c) and extract longitudinal affect aggregates from token probabilities. We evaluate our results against survey data of the

British adult population and find strong positive and significant correlations for a subset of collective emotions. A comparison of our method with a traditional classification model trained on labeled data produces strongly agreeing results. We demonstrate robustness against prompt variation with a survey instrument that measures the same affective phenomena (Clark and Watson 1994). Experiments with synthetically mixed data indicate the internal validity of our method. We exemplary show that our method can be used for the extraction of more complex collective attitudes as well.

Contribution The main contribution of this work is that it extends previous *in silico* surveys with LLMs to a *longitudi-nal* setting in which survey data and LLM training data are temporally aligned. Our method is less sensitive to strata-dependent biases embedded in pre-trained LLMs since we fine-tune LLMs on user-generated data and focus on longitudinal changes in affect aggregates. Compared to previous methods that extract affect aggregates from user-generated data, our approach is not fixed to a specific survey question and neither relies on expensively curated dictionaries, nor on labeled data. We provide a Python implementation under MIT license alongside our paper to facilitate replication and future work on extracting affect aggregates with LLMs¹.

Related Work

Attitudes, Opinions, and Values in LLMs

Many recent studies have investigated the application of longitudinal surveys to LLMs (e.g. Argyle et al. 2023; Bisbee et al. 2023; Atari et al. 2023; Von Der Heyde, Haensch, and Wenz 2023) Researchers typically prompted LLMs with survey questions from existing questionnaires that were originally developed to survey human populations (Ma et al. 2024; Agnew et al. 2024). This setup assumes that the diverse data on which LLMs were trained enables them to be viable proxies for population-level estimates of individual attributes. However, most research neglects the temporal alignment of training data and survey data, i.e., the fieldwork periods in which survey data was collected. We propose Temporal Adapters fine-tuned on user-generated data with a known date of creation to improve longitudinal alignment.

¹https://github.com/dess-mannheim/temporal-adapters



Figure 1: **Illustration of Temporal Adapters.** First, we gather weekly text data from a panel of Twitter users and fine-tune Temporal Adapters for Llama 3 8B with it. Then, we prompt the fine-tuned model with established survey questions, one week at a time, and extract affect aggregates from the answer options' token probabilities. Temporal Adapters enable longitudinal analyses of affect aggregates from social media data by temporally aligning LLMs.

Research has shown that openly available, pretrained LLMs produce estimates that can be both culturally and politically biased when compared to population-level survey data (e.g., Motoki, Pinho Neto, and Rodrigues 2023; Santurkar et al. 2023; Hartmann, Schwenzow, and Witte 2023; Adilazuarda et al. 2024). Since our method is based on finetuning an existing LLM, we likely inherit some of these biases. Still, by fine-tuning the model on user-generated data, we control the sampling process of whose attitudes, opinions, and values the model will learn.

LLMs were found to be sensitive to survey wording (Tjuatja et al. 2024; Dominguez-Olmedo, Hardt, and Mendler-Dünner 2023; Röttger et al. 2024; McIlroy-Young et al. 2024), and the particular answer scoring method (Wang et al. 2024b,a). We prompt each answer option separately to circumvent order effects and we use an answer-prefix to be less dependent on first-token probabilities. We demonstrate robustness on prompt wording by extracting collective emotions with multiple survey instruments.

Simulating Survey Participants with LLMs LLMs used for surveys are so far either studied as they were pre-trained, or they are fine-tuned on survey data (Ramezani and Xu 2023; Kim and Lee 2023). With pre-trained models, researchers typically use *silicon sampling* and prepend a diverse set of persona description, often derived from existing survey data, to the survey question (Argyle et al. 2023; Sun et al. 2024). A major drawback of silicon sampling is that is purely relies on a model's pre-training data, potentially misportraying parts of society (Wang, Morgenstern, and Dickerson 2024), and with no way of learning from the attitudes, opinions, and values that they voice. Our method is instead based around incorporating data generated by specific users at a specific time into LLMs used for surveys.

Kim and Lee (2023) added embeddings obtained from LLMs to a neural network to predict longitudinal changes in opinions. We instead work directly with a state-of-the-art LLM, to leverage its general language capabilities, and to extract affect aggregates using the same survey question that was asked to human survey participants. Ramezani and Xu (2023) trained an LLM on existing survey data and were able to improve similarity to human survey responses. The training of Temporal Adapters is independent of existing survey data and allows us to extract affect aggregates from specific user generated datasets.

Measuring Affect in Surveys

Our primary focus in this paper is on extracting emotion aggregates from user-generated social media data, even if our setup does not prescribe which survey question we use to extract collective affect, i.e., which construct we want to study. In social psychology and sociology, *affect* is used as the most general term that encompasses phenomena such as emotions and moods (Mohiyeddini and Bauer 2013; Rogers and Robinson 2014). *Emotions* are typically more specific and short-lived, lasting between minutes and days, while *moods* describe phenomena that are more diffuse and last between hours and weeks (Oatley and Johnson-Laird 2014), but the temporal distinction is not completely clear-cut.

Felt Emotions and Reactivity Researchers generally distinguish between *felt* and *expressed* emotions or moods (Rogers and Robinson 2014). Neither surveys nor social media data can tap into internal, felt emotions directly, so we focus only on expressed emotions in this paper. Survey measures of affect are further influenced by *recall bias* and by reactivity because of their retrospective nature. *Reactivity* refers to the fact that 'the very knowledge that one is

being observed can alter emotional experience' (Rogers and Robinson 2014, p. 286). Survey participants might also be more or less willing to report different emotions. Affect is an inherently social phenomena and central to interpersonal communication (Lively and Weed 2016), so social media platforms might intuitively seem like the perfect source of found data on expressed emotions. In contrast to surveys, posting on social media can be more immediate and less distanced from the affective experience itself. Working with found data is also considerably cheaper than surveying.

Social Media Affect Macroscopes

Text data generated by users on social media has been used to study election outcomes (Gayo-Avello 2013) and consumer confidence (Pasek et al. 2018), with challenging results. One of the most common applications is the extraction of *emotion macroscopes* (Golder and Macy 2011; Garcia et al. 2021; Pellert et al. 2022), in particular in response to catastrophic events (Garcia and Rimé 2019; Jones and Silver 2020) such as the beginning of the COVID-19 pandemic in 2020 (e.g., Valdez et al. 2020; Lwin et al. 2020; Ashokkumar and Pennebaker 2021; Metzler et al. 2023).

Similar to our setup, estimates are aggregated for specific groups, or on a population level. In contrast to previous approaches, our method does not rely on labeled training data or on expensively created dictionaries.

Sampling Biases and Performative Behavior Social media data suffers from sampling biases and performative behavior of users due to platform effects and community norms (Sen et al. 2021). Estimates obtained from social media samples often do not generalize to a target population because of differences in internet penetration and platform use (Amaya et al. 2020). In addition, behavioral differences between groups lead to biases in representation (Olteanu et al. 2019). In other words, different groups are more or less inclined to express emotions or attitudes on social media. Still, previous studies found strongly positive and robust correlations between collective emotions extracted from social media and survey responses on a population level (Garcia et al. 2021; Pellert et al. 2022). We extend this research by introducing an extraction method that is more flexible by extracting constructs directly through prompting LLMs with survey questions.

Neither self-reported affect in surveys, nor emotions extracted from user-generated social media data are by themselves perfectly accurate estimates. Since attitudes have a strong affective component (Bergman 1998), attitudes suffer from similar biases in both survey data and social media data. Our method, LLMs with Temporal Adapters, enables the flexible application of inference questions from the extensive, well-crafted collection of survey instruments for the extraction of longitudinal affect aggregates from social media data. Comparing population-level estimates obtained from both surveys and social media data will help us establish a more robust understanding of affective phenomena.

Longitudinal Datasets

We extract affect aggregates from Twitter panel data for Great Britain, covering 35 weeks from November 2019 to June 2020. This time frame includes both New Year 2020 as well as the first UK COVID-19 lockdown on March 23rd, 2020. It allows us to investigate both seasonal patterns and a catastrophic event that had a large impact on emotions and attitudes, as identified in previous research. We decided against a larger time frame due to the large amount of computational resources required to fine-tune our model.

Twitter Panel Data

We create a panel of in total 21,576 Twitter users (13.6 million tweets) with the following method: We use the commercial service Brandwatch to sample 10,000 tweets per day from accounts in Great Britain between the beginning of 2019 and March 2023. Brandwatch determines users' locations based on profile information and geo-tagged data (Brandwatch 2020b). Next, we identify individual users in that sample and removed accounts that were classified as organizations as well as accounts with too low or too high activity levels. We select 24,000 users at random, half of which Brandwatch classified as 'female' and half of which were classified as 'male' based on self-reported profile information (Brandwatch 2020a). We retrieve their full timelines of tweets including retweets and replies back to the beginning of 2019. On average, for the 35 weeks from November 2019 to June 2020, we have 385,000 tweets per week available.

Questionnaires and Survey Data

We extract affect aggregates by querying an LLM with existing survey questionnaires and we compare our results to publicly available survey data that was collected using the same questions. The survey data was collected from a British online panel with a target population of all British adults aged 18+ (YouGov 2024b). The panel provider uses active sampling and post-survey adjustment weights on age, gender, social class, region, and level of education.

Britain's Mood, Measured Weekly To extract collective emotions, we use the question developed by YouGov (2024a):

Broadly speaking, which of the following best describe your mood and/or how you have felt in the past week?

This question uses a multi-code answer scale, i.e., multiple of the following answer options can be selected: 'happy', 'sad', 'energetic', 'apathetic', 'inspired', 'frustrated', 'optimistic', 'stressed', 'content', 'bored', 'lonely', and 'scared'. We exclude the instructions on how to answer as well as the answer options 'other' and 'don't know' from our setup, as these are tailored to human survey response behavior. We assess our results by comparing them with aggregate survey data gathered from the above described British panel on the same question. Survey data is available in weekly waves as aggregates over 1890 to 2081 participants per wave. **PANAS-X** We further investigate collective emotions using the extended version of the Positive and Negative Affect Schedule (PANAS-X) developed by Clark and Watson (1994). This survey instrument asks participants to indicate to what extend they feel they have felt like a series of carefully compiled adjectives. It comes with a series of time instructions, ranging from how participants feel in the moment to in general. The instrument has been validated for *state affect*, i.e., short-term fluctuations in mood, as well (Clark and Watson 1994). We create a prompt based on the 'week' instruction, to measure emotions in the same time frame as in YouGov's wording, as follows:

To what extend have you felt [adjective] during the past week?

We focus on the two emotions included in both the PANAS-X and in YouGov's survey data, *scared* and *sad*. The answer options are 'very slightly or not at all', 'a little', 'moderately', 'quite a bit', and 'extremely', to be answered for each adjective – in contrast to YouGov's answer options, which are directly multiple choice among emotions. We extract answer probabilities for each adjective and each answer option, and combine all of them into a single score for each emotion, according to the instructions provided by Clark and Watson (1994). We compare our results to survey data gathered by YouGov in *Britain's Mood, Measured Weekly* for the respective emotions.

The National Health Service (NHS) In addition to collective emotions, we also extract a collective attitude towards the NHS with the following question (YouGov 2024c):

Do you expect the National Health Service to get better, worse or stay the same over the next few years?

We spell out the abbreviation 'NHS' since we query for this survey question without additional context. We include the answer options 'get better' and 'get worse' into our analysis and compare our results to survey data from YouGov on the same question. Survey data is available in monthly waves as aggregates over 1618 to 1817 participants per wave.

Temporal Adapters for LLMs

Figure 1 provides an overview of our proposed setup, comprising two separate steps: First, we fine-tune Temporal Adapters for Llama 3 8B on the Twitter panel data described in the previous section. Second, we prompt the model with one of the selected survey questions and one weekly adapter activated at a time. We extract token probabilities for each answer option across all weeks, which we combine into longitudinal macroscopes of mood and attitudes.

To validate our results, we cross-correlate them with the respective survey data. We also include results from a BERT-based emotion detection model trained on labeled data (Camacho-Collados et al. 2022) for comparison. Finally, we conduct experiments on synthetically mixed, labeled data to demonstrate the internal validity of our attitude extraction method.



Figure 2: **Temporal Adapter Fine-Tuning.** We concatenate each week's tweets into chunks for batch-based finetuning. While fine-tuning each week's parameter-efficient LoRA (Hu et al. 2021) adapter, the original model weights are kept frozen. Fine-Tuning is performed with the causal language modeling task, i.e., next-token prediction.

Temporal Adapter Fine-Tuning

Following the wording of YouGov's *Britain's Mood, Measured Weekly*, we split our social media panel data into weekly subsets, containing seven days of text data leading up to the next YouGov survey wave. We obtain on average 385,000 tweets per training set, but the amount varies seasonally and around major events like the UK COVID-19 lockdown. We concatenate each training set into sequences of 512 tokens to facilitate batch training.

We select the base pretrained version (i.e., not instruction tuned) of a high-performing and openly available model, Llama 3 8B (AI@Meta 2024), to fine-tune the model on plain text data. We fine-tune Temporal Adapters with the causal language modeling objective, i.e. predicting next tokens, as shown in Figure 2. Adapters are a parameter-efficient approach to fine-tuning LLMs that allows researchers to easily swap model properties (Pfeiffer et al. 2023). LoRA adapters (Hu et al. 2021) add rankdecomposition matrices to each layer of the transformer architecture. When fine-tuning an LLM with LoRA adapters, all original model weights are frozen and only the added weights are trained. We train LoRA adapters of rank 128 as a compromise between fine-tuning efficiency and training quality. This is twice the amount of parameters as the authors of the original LoRA paper considered necessary (Hu et al. 2021), but still only 0.67% of the weights that would be trained when fully fine-tuning the model.

We conduct a partial hyperparameter search by training adapters of varying rank, with different learning rates, and for up to 8 epochs. We find that fine-tuning adapters works best for our purposes with a small learning rate of $5 * 10^{-6}$ and at most 1 training epoch. The training loss is mostly stable after 0.5 epochs at an average of 3.4 across all adapters. In a preliminary experiment with a larger learning rate of 10^{-4} , training and test set loss decreased to around 2.6 after 8 training epochs but the cross-correlations of the attitudes we extracted with survey data were much lower. This is most likely due to the model over-fitting on the training data (tweets) that is quite different from the survey questions we prompt at inference time. We train adapters with 3 different training seeds for each week to better understand the reliability of our training method. We used bfloat16 and the adamw_torch_fused optimizer with a batch size of 6 and 4 gradient accumulation steps. The training was conducted on two NVIDIA H100 GPUs in Distributed-Data-Parallel mode and takes approximately 20 minutes per adapter.

Extracting Survey Answers

Once we have finished adapter training, we focus on model inference and extract answers to survey questions as follows. We first concatenate a survey question with each of its *n* answer options, obtaining *n* separate prompts. We experiment with an optional answer prefix that is build for the particular survey question. For instance, for *Britain's Mood, Measured Weekly*, we add the optional answer prefix 'I felt' which is followed by the answer options 'happy', 'sad' etc. This is to accommodate for the fact that first-token probabilities can be unreliably for scoring survey answers from LLMs (Wang et al. 2024b). We prompt each answer option separately and without additional labels as previous research has shown order effects and the tendency for LLMs to prefer certain answer labels (Tjuatja et al. 2024; Dominguez-Olmedo, Hardt, and Mendler-Dünner 2023; Wang et al. 2024b).

Based on a survey answer scoring method used in previous research (Wu et al. 2024; Naous et al. 2024), we perform a single forward pass of the concatenated question, optional answer prefix, and answer option through the LLM. We gather the token probabilities from the last LLM layer for each token in the answer option after applying softmax. If an answer option consists of multiple tokens, we multiply their probabilities. We further vary *temperature*, a normalization parameter applied to the final softmax function, between 0.25 and 4. Our extraction method is deterministic, i.e., if the weights of an LLM are kept constant, so will be the answer probabilities we extract. Unless otherwise noted, we report results for temperature 1, i.e., standard softmax.

We perform survey answer extraction with each weekly adapter separately activated in the LLM and for each of the 3 seeds. Since swapping adapters and inference is computationally inexpensive, we extract survey answers after training each adapter for 1 epoch, as well as after every 50 steps of training. We obtain a time series for each combination of question, answer options, and set of inference hyperparameters. YouGov presents weekly estimates using a smoothed trend line to reduce random fluctuations due to sampling variability. We follow this approach and apply a 3 week rolling average. For plotting time series, we apply min-max normalization and show the mean probability for each extracted survey answer across all 3 seeds.

Comparison and Validation

Cross-Correlating Survey Data We evaluate our longitudinal macroscopes of mood and attitudes by crosscorrelating them with the respective survey data using Pearson's correlation coefficient. YouGov publishes weekly survey results on *Britain's Mood*, which results in 35 data points in our observation period that we can use to compute cross-correlation. Results on British attitude towards the NHS are only published monthly, i.e., we are left with 9 relevant data points. We perform permutation tests with 10.000 permutations of our time series, while keeping the survey data intact, to obtain a significance value for our cross-correlations.

Baseline Comparison Model We compare our results to emotion aggregates from a BERT-based model for emotion detection, TweetNLP (Camacho-Collados et al. 2022). This baseline model will help us differentiate between issues in our training data, e.g. coverage error, or differences in expression of emotion, and measurement error specific to our research setup. TeetNLP was pre-trained on the tweet_eval dataset that consists of tweets labeled with 11 emotions (Mohammad et al. 2018). The model achieves a macro F1 score of 0.72 on the tweet_eval test set. We classify each tweet in our weekly training datasets separately and calculate the share of every emotion in every week. Being pre-trained on Twitter data makes TweetNLP particularly suited as a baseline model for our setup. However, only 4 of the emotions classified by TweetNLP match with emotions included in Britain's Mood, Measured Weekly: 'fear'/'scared', 'sadness'/'sad', 'joy'/'happy', and 'optimism'/'optimistic'. We again cross-correlate the results obtained from TweetNLP with the survey data and perform a permutation test.

Synthetically Mixed Data To explore the internal validity of our mood macroscopes, we perform additional experiments with synthetically mixed data that is labeled as 'happy' or 'sad'. With this setup, we can investigate whether a different prevalence of emotion signals indeed corresponds to different affect aggregates obtained from our method. We hypothesize a linear relationship between the amount of 'sad' tweets in the training data and the token probability for 'sad' as an answer option, and vice versa for 'happy'. There are 1163 tweets labeled 'happy' and 1326 tweets labeled 'sad' in the tweet_eval dataset. From these tweets, we select 1163 using 11 splits: 100% happy + 0% sad, 90%happy + 10% sad, ..., and 0% happy + 100% sad. We then train adapters on each of these splits using the same hyperparameters as described above. We repeat the splitting and training procedure using 10 random seeds to improve robustness. Since the synthetically mixed training datasets are much smaller than our weekly Twitter datasets, the number of training steps/epochs is not directly comparable.



Figure 3: Affect Aggregates Extracted from Temporal Adapters. We extract answer probabilities by prompting a weekly fine-tuned Llama 3 8B with the same question wording as in the survey (YouGov 2024a), and compare them to the respective weekly survey data. The time series are min-max normalized and a 3 week rolling average is applied. The shaded orange area indicates minimum and maximum LLM answer probabilities across 3 training seeds. Our results descriptively show in the plot a similar trend of both signals and we find strong positive and significant (p < 0.01) cross-correlation between LLM probabilities and the survey data. Additional time series are provided in Figures 7 and 8 in the Appendix.

Results

Extracted Emotions Highly Correlate with Survey Data Figure 3 shows time series for two collective emotions, scared and happy, that we extract using weekly Temporal Adapters. We use the original survey question that was developed by YouGov (2024a) and compare our results to nationally representative survey data. We cap the training steps such that we have the same amount of training data across weeks. The plot shows the mean answer probability across 3 training seeds, with minimum and maximum probabilities, after a 3 week rolling average was applied and after the time series were min-max normalized. Time series for other emotions are provided in Figures 7 and 8 in the Appendix.

Overall, we observe similar trends between our estimates of collective emotion and the survey data. Scared spikes on the March 23rd, 2020, when the first UK lockdown in response to the COVID-19 pandemic was announced². Another increase can be seen in our estimate of collective fear around June 1st 2020, at a time when lockdown restrictions were partially lifted³. Fear also sees an increase around Christmas 2019. The survey data similarly indicates the largest level of collective fear around the announcement of the lockdown and a small increase around Christmas, but the second spike around June 1st, 2020 is not observed in the survey. This is likely due to sampling errors in the social media data, i.e., more concerned people being active on Twitter, or due to differences in the expression of emotions on social media and in surveys. TweetNLP shows the same second spike in 2020 when used to classify the same training data, as shown in Appendix Figure 8.

For collective *happiness*, we similarly extract estimates with good face validity. Happiness decreases rapidly in the two weeks before the nationwide lockdown was announced and recovers after another two weeks, presumably when people acclimatized with the new situation. This follows the trend that we observe in survey data. Again, our estimate deviates from survey data around June 1st 2020, when restrictions were partially lifted.

Compared to TweetNLP, which extracts the lowest happiness after June 1st, however, our method's estimate is closer to the YouGov survey data in that it identifies the lowest happiness around March 23rd. It should be addressed that we observe larger confidence intervals across 3 training seeds, and larger week-by-week fluctuation for 'happy' as compared to 'scared'. One possible explanation for this would be that there is less information about happiness in our training data, which would lead to less numerical influence in adapter fine-tuning, and thus result in larger confidence intervals after min-max scaling. Another explanation could be that information about happiness is less evenly distributed in our training data. We sample an equal amount of text for each week and randomly shuffle the training data, so less even distribution of information about happiness would increase the relevance of different training seeds. Thirdly, the LLM could be less capable of learning information about happiness during training, leading to more error surrounding the concept.

Taking a step back, we calculate cross-correlations (Pearson) between the longitudinal data we extract and the respective survey data, and perform a significance test with 10,000 permutations. Figure 4 shows the results for all emotions found in *Britain's Mood, Measured Weekly* across 3 training seeds, as well as the cross-correlation for our baseline method, TweetNLP (Camacho-Collados et al. 2022). We observe three groups of emotions: (i) emotions that show a strong positive, significant correlation, (ii) emotions that

²https://www.theguardian.com/world/2020/mar/23/borisjohnson-orders-uk-lockdown-to-be-enforced-by-police, Accessed Sept 8th, 2024

³ for England, see for instance https://www.legislation.gov.uk/u ksi/2020/558/made, Accessed Sept 8th, 2024

are weakly positively or negatively correlated, and (iii) optimism, which is strongly negatively correlated. In the first group we find both positively and negatively connoted emotions, and in particular affective phenomena that are sometimes referred to as being part of a set of basic emotions: 'happiness', 'fear', and 'sadness' (Ekman 1992). For collective emotions where an estimate can be extracted from TweetNLP, i.e., where labeled training data was available, we find that our method produces comparable results. We notably also obtain strong positive, significant correlations for emotions for which no TweetNLP estimate is available: 'frustration', 'boredom', and 'being energetic'. For group (ii), is remains unclear why these collective emotions are hard to estimate from the data that we have available. One possible explanation might be that in contrast to the affective phenomena found in group (i), group (ii) contains more phenomena that clearly fall under 'mood' rather than 'emotion', i.e., they are more diffuse and longer lived, which might contradict typical social media behavior. Finally, we find a significant strong negative cross-correlation for 'optimistic', both for our estimate as well as for the baseline model. This points to general differences in the expression of this emotions between surveys and social media.

Experiments with Synthetically Mixed Data We create synthetically mixed training data from the tweet_eval dataset (Mohammad et al. 2018) in 11 splits ranging from 100% sad to 100% happy. We then prompt the model with the same question as used in Britain's Mood, Measured Weekly and extract answer probabilities for the answer options 'happy' and 'sad'. We hypothesize a positive linear correlation between the share of 'happy' tweets in the training data and the probability for answer option 'happy', and vice versa for 'sad'. Figure 5 shows the mean and standard deviation of extracted answers across 10 training seeds. We find strong positive correlations for both answer options, which supports the internal validity of our method - more 'happy' tweets actually lead to a higher answer probability for 'happy' and vice versa for 'sad'. For both emotions, the relationship between training data and extracted answer is surprisingly linear. Our answer scoring method does not ensure that the probabilities across all answer options add up to 1, since we apply softmax on the last layer of the LLM, and each answer probability depends on probabilities of all other tokens in the vocabulary. The wide error bands indicate large random error due in training and underline the importance of evaluating results obtained from multiple seeds.

In preliminary experiments, we investigated different hyperparameters for model training and for the extraction of affect aggregates, see Figure 11 in the Appendix. We investigated cross-correlations after different numbers of training steps and epochs and found little benefit in fine-tuning the model for more than 1 epoch. While longer fine-tuning further improves training loss in the causal language modeling objective, the highest cross-correlations with survey data are observed relatively early on. This is likely due to the language model overfitting on the training data, which itself is quite different from the questions that we use for answer extraction. Similar also applies to learning rate – we





Figure 4: Several Extracted Emotions Highly Correlate with Survey Data. We cross-correlate the answers we extract from Llama 3 with the respective British survey data (YouGov 2024a). Across 3 training seeds, we show minimum and maximum correlation with error bars and indicate the worst p value (*p < 0.05,**p < 0.01,***p < 0.001). Our results vary strongly between emotions. They are in line with the baseline model's estimates, and extend them to additional emotions by not requiring labeled training data.

experimented with learning rates between 10^{-3} and 10^{-6} and found that smaller learning rates generally perform better. Only 10^{-6} was too small, as the training loss remained almost constant, so we opted for $5 * 10^{-6}$ instead. Finally, we found that training all adapters with the same amount of training data generally works better than training all adapters for 1 epoch on all the data that was available in this week. In other words, we extracted answers for all Figures presented previously in this section after training each adapter for 350 steps instead of 1 epoch. For answer extraction, we investigated the effect of temperature and of whether or not an answer prefix was attached to the answer options. A lower temperature leads to less evenly distributed token probabilities, similar to 'enhancing contrast'. We found that this is beneficial when extracted attitudes have low noise, but that it can also increase random fluctuation. We also found that lower temperatures tend to overemphasize some answer options. Finally, adding an answer prefix lead to more consistent results. This is in line with previous findings on firsttoken probabilities (Wang et al. 2024b) and is likely related to LLMs being trained to create text rather than to answer surveys with only an answer option.



Figure 5: Internal Validity Demonstrated in Experiments with Synthetically Mixed Data. We synthetically mix LLM training data with splits ranging from data that is labeled 100% sad to 100% happy. We then extract answers to YouGov (2024a)'s survey question at each split, and show mean and standard deviation over 10 training seeds. The results support the internal validity of our extraction method, but also highlight random error in training and a nonlinear relationship between training data ratio and extracted estimate.

Robustness Across Survey Instruments We implement two additional survey questions besides the one used in Britain's Mood, Measured Weekly. First, we focus on demonstrating the robustness of our extraction method and extract the collective emotions scared and sad the PANAS-X survey instrument. The instrument works quite differently from YouGov's questionnaire in that it measures each emotion with multiple adjectives, with the same answer options for each adjective: 'very slightly or not at all', 'a little', etc. Again, we cross-correlate the extracted collective emotions with the YouGov survey data - Figure 6a shows the results from 3 training seeds across the different extraction methods. We find that both scared and sad achieve crosscorrelations comparable to the ones we extracted with the original YouGov question wording. This supports the robustness of our extraction method across prompt formulations, i.e., survey instruments, when measuring the same construct.

Application to Attitude Aggregates Second, we go beyond measuring collective emotions and tested a question that measures attitudes towards the National Health Service (NHS). YouGov started running their weekly and monthly surveys on different days in May 2020, so we train 3 additional adapters to extract attitudes in exactly the same weeks that the monthly survey was run. YouGov runs most of its 'trackers' only in monthly waves, including on the attitude towards the NHS. This means that during the time period that we study, we only have 9 data points available. Here we see a potential clear benefit of extracting attitudes from social media data – it is much less expensive and can have a higher temporal resolution. Figure 6b shows the crosscorrelation of the two available answer options for the 9 available data points between our estimates and the survey data – for time series, see Figure 10 in the Appendix. While we find positive correlations, they are not significant across all random seeds and the different seeds create a large confidence interval. Reasons for this might be measurement error in our extraction method for a more complex construct, and less discussion around this particular topic in our training data. Still, this result shows that our method can, in principle, be applied to the extraction of more complex attitudes as well.

Discussion

Limitations Our method focuses on the extraction of separate time series data, rather than comparing the prevalence of moods or attitudes cross-sectionally. The reason is that the relative probabilities of answer options in a crosssectional sense are strongly influenced by LLM pre-training. Future research should combine these, so far orthogonal, approaches by correcting for differences that stem from model pre-training. The results we present were obtained by training weekly Temporal Adapters, following exactly the intervals in which the respective survey waves were conducted. This allowed for improved comparability between our estimates and survey data, but higher temporal resolution, a potential benefit of using social media data, should be investigated in future research. In regions where survey data is available, survey data could in the future be used as a prior for predicting affect aggregates with higher temporal resolution from social media data. Yet, our method remains applicable to regions in which no (recent) survey data is available, e.g., because of lack of infrastructure or an ongoing violent conflict. Like any other attitude extraction method, our approach requires validation in additional contexts, with possible design decisions ranging from model training hyperparameters to the exact question wording and the answer extraction method. Future research should further investigate the applicability of this method to subgroup analysis, and in particular to whether affect aggregates can be extracted with the same quality for different subpopulations, even if sampling error is overcome. To facilitate such efforts, we publish our Python code under MIT license⁴.

Ethical Considerations YouGov pays its panel members adequately for their participation in surveys and is a member of several market research organisations (British Polling Council, ESOMAR, MRS) whose standards it adheres to (YouGov 2024b). We work with social media data that was publicly available at the time that we gathered it. We do not publish raw text data, or the Temporal Adapters that we trained, to protect the privacy of the individuals whose tweets are included in our training data. We do, however, publish the code that is needed to replicate our results under MIT license to support future research. Our method relies on more training data than the vast majority of social media users produce in a week, so it only works on an aggregate

⁴https://github.com/dess-mannheim/temporal-adapters



(a) Robustness Across Multiple Survey Instruments: YouGov & PANAS-X

(b) Application to Attitude Aggregates

Figure 6: We extract the same collective emotions with (a) an additional survey instrument (PANAS-X, Clark and Watson 1994) and (b) separately extract a collective attitude towards the NHS. We cross-correlate our results with the respective YouGov survey data. We find that our method is robust across multiple survey instruments and can be applied to the extraction of collective attitudes as well. Across 3 training seeds, we indicate minimum and maximum correlation with error bars and show the worst p value (*p < 0.05,**p < 0.01,***p < 0.001). For time series, see Figures 9 and 10 in the Appendix.

level. This limits the risks associated with profiling individuals through our method, even if we cannot fully rule out unethical applications. Our main results (Figure 4) highlight the importance of validation and comparison to survey data for each context and construct of interest. Still, an application of our method without such rigor could inform wrong or discriminatory downstream decisions, either because of bias in training data or because of measurement error.

Conclusion

This paper expands the inventory of methods for affect analysis available to our research community. Our work closes a temporal misalignment gap in previous surveys with LLMs by proposing Temporal Adapters that are trained on longitudinal social media data. To the best of our knowledge, our paper is the first to extend the analysis of affect in LLMs to a *longitudinal* setting. Temporal Adapters for LLMs open up new ways for studying affect aggregates in social media data longitudinally. More broadly, they may also enable future, more temporally-aligned LLM-based studies of human attitudes, values, and opinions.

References

Adilazuarda, M. F.; Mukherjee, S.; Lavania, P.; Singh, S.; Dwivedi, A.; Aji, A. F.; O'Neill, J.; Modi, A.; and Choudhury, M. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. ArXiv:2403.15412 [cs].

Agnew, W.; Bergman, A. S.; Chien, J.; Díaz, M.; El-Sayed, S.; Pittman, J.; Mohamed, S.; and McKee, K. R. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–12. ArXiv:2401.08572 [cs].

AI@Meta. 2024. Llama 3 Model Card. https://github.com/metallama/llama3/blob/main/MODEL_CARD.md. Accessed: 2024-09-02.

Amaya, A.; Bach, R.; Kreuter, F.; and Keusch, F. 2020. Measuring the Strength of Attitudes in Social Media Data. In Hill, C. A.; Biemer, P. P.; Buskirk, T. D.; Japec, L.; Kirchner, A.; Kolenikov, S.; and Lyberg, L. E., eds., *Big Data Meets Survey Science*, 163– 192. Wiley, 1 edition. Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Ashokkumar, A.; and Pennebaker, J. W. 2021. Social media conversations reveal large psychological shifts caused by COVID-19's onset across U.S. cities. *Science Advances*, 7(39): eabg7843.

Atari, M.; Xue, M. J.; Park, P. S.; Blasi, D. E.; and Henrich, J. 2023. Which Humans? preprint, PsyArXiv.

Bergman, M. M. 1998. A Theoretical Note on the Differences Between Attitudes, Opinions, and Values. *Swiss Political Science Review*, 4(2): 81–93.

Bisbee, J.; Clinton, J.; Dorff, C.; Kenkel, B.; and Larson, J. 2023. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. preprint, SocArXiv.

Brandwatch. 2020a. Forsight: User Guide. https://web.archive.or g/web/20240513122948/https://www.brandwatch.com/wp-conten t/uploads/2020/10/Crimson-Hexagon-ForSight-User-Guide.pdf. Accessed: 2024-05-13.

Brandwatch. 2020b. Location Methodology. https://web.archive.or g/web/20210527091937/https://www.brandwatch.com/wp-conten t/uploads/2020/10/CrimsonHexagon_Location_Methodology.pdf. Accessed: 2021-05-27.

Camacho-Collados, J.; Rezaee, K.; Riahi, T.; Ushio, A.; Loureiro, D.; Antypas, D.; Boisson, J.; Espinosa-Anke, L.; Liu, F.; Martínez-Cámara, E.; Medina, G.; Buhrmann, T.; Neves, L.; and Barbieri, F. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. ArXiv:2206.14774 [cs].

Clark, L. A.; and Watson, D. 1994. The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form. Institution: University of Iowa.

Dominguez-Olmedo, R.; Hardt, M.; and Mendler-Dünner, C. 2023. Questioning the Survey Responses of Large Language Models. ArXiv:2306.07951 [cs].

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. ArXiv:2306.16388 [cs].

Ekman, P. 1992. Are there basic emotions? *Psychological Review*, 99(3): 550–553.

FORCE11. 2020. The FAIR Data principles. https://force11.org/ info/the-fair-data-principles/.

Garcia, D.; Pellert, M.; Lasser, J.; and Metzler, H. 2021. Social media emotion macroscopes reflect emotional experiences in society at large. ArXiv:2107.13236 [cs].

Garcia, D.; and Rimé, B. 2019. Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack. *Psychological Science*, 30(4): 617–628.

Gayo-Avello, D. 2013. A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, 31(6): 649–679.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Golder, S. A.; and Macy, M. W. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333(6051): 1878–1881.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. ArXiv:2106.09685 [cs].

Jones, N. M.; and Silver, R. C. 2020. This is not a drill: Anxiety on Twitter following the 2018 Hawaii false missile alert. *American Psychologist*, 75(5): 683–693.

Kim, J.; and Lee, B. 2023. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. ArXiv:2305.09620 [cs].

Lively, K. J.; and Weed, E. A. 2016. The Sociology of Emotions. In Feldman Barret, L.; Lewis, M.; and Haviland-Jones, J. M., eds., *Handbook of Emotions*, 66–81. New York: Guilford Press, fourth edition.

Lwin, M. O.; Lu, J.; Sheldenkar, A.; Schulz, P. J.; Shin, W.; Gupta, R.; and Yang, Y. 2020. Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends. *JMIR Public Health and Surveillance*, 6(2): e19447.

Ma, B.; Wang, X.; Hu, T.; Haensch, A.-C.; Hedderich, M. A.; Plank, B.; and Kreuter, F. 2024. The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. ArXiv:2406.11096 [cs].

McIlroy-Young, R.; Brown, K.; Olson, C.; Zhang, L.; and Dwork, C. 2024. Set-Based Prompting: Provably Solving the Language Model Order Dependency Problem. ArXiv:2406.06581 [cs].

Metzler, H.; Rimé, B.; Pellert, M.; Niederkrotenthaler, T.; Di Natale, A.; and Garcia, D. 2023. Collective emotions during the COVID-19 outbreak. *Emotion*, 23(3): 844–858.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 1–17. New Orleans, Louisiana: Association for Computational Linguistics.

Mohiyeddini, C.; and Bauer, S. 2013. What is an Emotion? In Mohiyeddini, C.; Eysenck, M. W.; and Bauer, S., eds., *Handbook of Psychology of Emotions*, volume Volume 1: Recent Theoretical Perspectives and Novel Empirical Findings, 3–10. New York: Nova Publ.

Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2023. More human than human: measuring ChatGPT political bias. *Public Choice*, 198: 3–23.

Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. ArXiv:2305.14456 [cs].

Oatley, K.; and Johnson-Laird, P. 2014. Cognitive approaches to emotions. *Trends in Cognitive Sciences*, 18(3): 134–140.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2: 13.

Pasek, J.; Yan, H. Y.; Conrad, F. G.; Newport, F.; and Marken, S. 2018. The Stability of Economic Correlations over Time. *Public Opinion Quarterly*, 82(3): 470–492.

Pellert, M.; Metzler, H.; Matzenberger, M.; and Garcia, D. 2022. Validating daily social media macroscopes of emotions. *Scientific Reports*, 12(1): 11236.

Pfeiffer, J.; Ruder, S.; Vulić, I.; and Ponti, E. M. 2023. Modular Deep Learning. ArXiv:2302.11529 [cs].

Ramezani, A.; and Xu, Y. 2023. Knowledge of cultural moral norms in large language models. ArXiv:2306.01857 [cs].

Rogers, K. B.; and Robinson, D. T. 2014. Measuring Affect and Emotions. In Stets, J. E.; and Turner, J. H., eds., *Handbook of the Sociology of Emotions: Volume II*, 283–303. Dordrecht: Springer Netherlands. Series Title: Handbooks of Sociology and Social Research.

Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H. R.; Schütze, H.; and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. ArXiv:2402.16786 [cs].

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? ArXiv:2303.17548 [cs].

Sen, I.; Flöck, F.; Weller, K.; Weiß, B.; and Wagner, C. 2021. A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*, 85(S1): 399–422.

Sun, S.; Lee, E.; Nan, D.; Zhao, X.; Lee, W.; Jansen, B. J.; and Kim, J. H. 2024. Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information. Version Number: 1.

Tjuatja, L.; Chen, V.; Wu, S. T.; Talwalkar, A.; and Neubig, G. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. ArXiv:2311.04076 [cs].

Valdez, D.; Ten Thij, M.; Bathina, K.; Rutter, L. A.; and Bollen, J. 2020. Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *Journal of Medical Internet Research*, 22(12): e21418.

Von Der Heyde, L.; Haensch, A.-C.; and Wenz, A. 2023. Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion. preprint, SocArXiv.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2024. Large language models cannot replace human participants because they cannot portray identity groups. ArXiv:2402.01908 [cs].

Wang, X.; Hu, C.; Ma, B.; Röttger, P.; and Plank, B. 2024a. Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think. ArXiv:2404.08382 [cs].

Wang, X.; Ma, B.; Hu, C.; Weber-Genzel, L.; Röttger, P.; Kreuter, F.; Hovy, D.; and Plank, B. 2024b. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. ArXiv:2402.14499 [cs].

Wu, Q.; Khan, M. A.; Das, S.; Nanda, V.; Ghosh, B.; Kolling, C.; Speicher, T.; Bindschaedler, L.; Gummadi, K. P.; and Terzi, E. 2024. Towards Reliable Latent Knowledge Estimation in LLMs:

In-Context Learning vs. Prompting Based Factual Knowledge Extraction. ArXiv:2404.12957 [cs].

YouGov. 2024a. Britain's Mood, Measured Weekly. https://youg ov.co.uk/topics/politics/trackers/britains-mood-measured-weekly. Accessed: 2024-09-02.

YouGov. 2024b. Panel Methodology. https://yougov.co.uk/about /panel-methodology. Accessed: 2024-09-02.

YouGov. 2024c. Will the NHS get better or worse? https://yougov .co.uk/topics/politics/trackers/is-the-nhs-getting-better-or-worse. Accessed: 2024-09-02.

Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socioeconomic divide, or implying disrespect to societies or cultures? Yes, see Ethical Considerations
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Comparison and Validation
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Social Media Macroscopes
- (e) Did you describe the limitations of your work? Yes, see Limitations
- (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethical Considerations
- (g) Did you discuss any potential misuse of your work? Yes, see Ethical Considerations
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethical Considerations
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
- 2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? NA
- (b) Have you provided justifications for all theoretical results? NA
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
- (e) Did you address potential biases or limitations in your theoretical framework? NA
- (f) Have you related your theoretical results to the existing literature in social science? NA
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
- 3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? NA

- (b) Did you include complete proofs of all theoretical results? $\ensuremath{\operatorname{NA}}$
- 4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we publish the code under MIT license
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Comparison and Validation
- (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Ethical Considerations
- Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, without compromising anonymity...
 - (a) If your work uses existing assets, did you cite the creators? Yes, see Longitudinal Datasets
- (b) Did you mention the license of the assets? Yes
- (c) Did you include any new assets in the supplemental material or as a URL? No, to protect individual privacy
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethical Considerations
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Ethical Considerations
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
- 6. Additionally, if you used crowdsourcing or conducted research with human subjects, without compromising anonymity...
 - (a) Did you include the full text of instructions given to participants and screenshots? Yes, see Questionnaires and Survey Data
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, because survey data was gathered by YouGov
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No. YouGov does not publish detailed enough information, but adheres to ESOMAR/MRS standards.
- (d) Did you discuss how data is stored, shared, and deidentified? Yes, see Ethical Considerations

Appendix



Figure 7: **Time Series of Emotions from Temporal Adapters, Part 1.** We extract answer probabilities by prompting a weekly fine-tuned Llama 3 8B with the same question wording as in the survey (YouGov 2024a), and compare them to the respective weekly survey data. The time series are min-max normalized and a 3 week rolling average is applied. The shaded orange area indicates minimum and maximum LLM answer probabilities across 3 training seeds.



Figure 8: **Time Series of Emotions from Temporal Adapters, Part 2.** We extract answer probabilities by prompting a weekly fine-tuned Llama 3 8B with the same question wording as in the survey (YouGov 2024a), and compare them to the respective weekly survey data. The time series are min-max normalized and a 3 week rolling average is applied. The shaded orange area indicates minimum and maximum LLM answer probabilities across 3 training seeds.



Figure 9: **Time Series of Emotions from Temporal Adapters, Extracted with the PANAS-X Instructions.** We extract answer probabilities by prompting a weekly fine-tuned Llama 3 8B with a question based on the 'week' instructions of the PANAS-X inventory (Clark and Watson 1994). We combine the responses into a single score as designed by Clark and Watson (1994), and compare our results to the respective weekly survey data from (YouGov 2024a). The time series are min-max normalized and a 3 week rolling average is applied. The shaded orange area indicates minimum and maximum LLM answer probabilities across 3 training seeds.



Figure 10: **Time Series of Attitudes towards the NHS from Temporal Adapters.** We extract answer probabilities by prompting a weekly fine-tuned Llama 3 8B with the same question wording as in the survey (YouGov 2024c), and compare them to the respective weekly survey data. The time series are min-max normalized and a 3 week rolling average is applied. The shaded orange area indicates minimum and maximum LLM answer probabilities across 3 training seeds. YouGov survey data for comparison is only available in monthly waves.



Figure 11: **Results from Partial Hyperparameter Search.** We synthetically mix LLM training data with splits ranging from data that is labeled 100% sad to 100% happy. We then extract answers to YouGov (2024a)'s survey question at each split, and show mean and standard deviation over 10 training seeds. Unless otherwise noted, each plot shows results after 50 training steps, with learning rate $5 * 10^{-6}$, temperature 1, and using an answer prefix for answer extraction. We aim at a linear relationship between training mix and extracted answers, with low random error across training seeds.