Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models

Hongli Zhan[1] **Desmond C. Ong**[2] **Junyi Jessy Li**[1]
[1]Department of Linguistics, The University of Texas at Austin
[2]Department of Psychology, The University of Texas at Austin
{honglizhan, desmond.ong, jessy}@utexas.edu

Abstract

The emotions we experience involve complex processes; besides physiological aspects, research in psychology has studied *cognitive appraisals* where people assess their situations subjectively, according to their own values (Scherer, 2005). Thus, the same situation can often result in different emotional experiences. While the *detection* of emotion is a well-established task, there is very limited work so far on the automatic prediction of cognitive appraisals. This work fills the gap by presenting COVIDET-APPRAISALS, the most comprehensive dataset to-date that assesses 24 appraisal dimensions, each with a natural language rationale, across 241 Reddit posts. COVIDET-APPRAISALS presents an ideal testbed to evaluate the ability of large language models — excelling at a wide range of NLP tasks — to automatically assess and explain cognitive appraisals. We found that while the best models are performant, open-sourced LLMs fall short at this task, presenting a new challenge in the future development of emotionally intelligent models. *We release our dataset at* https://github.com/honglizhan/CovidET-Appraisals-Public.

1 Introduction

Emotions constitute a crucial aspect of people's lives, and understanding them has a profound impact on improving public mental health problems as well as policy-making (Choudhury and De, 2014; Gjurković and Šnajder, 2018; Arora et al., 2021; Uban et al., 2021). The emotions we experience involve complex processes: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations. These are called *cognitive appraisals*, and have been extensively studied in psychology through theoretical, behavioral, and hand-coded studies (Arnold, 1960; Lazarus, 1966; Lazarus et al., 1980; Roseman, 1984; Scherer et al., 1984; Smith and Ellsworth, 1985; Weiner, 1985; Clore and Ortony, 2000; Roseman and Smith, 2001; Scherer et al., 2001; Ellsworth and Scherer, 2003; Sander et al., 2005; Ong et al., 2015, 2019; Ortony et al., 2022; Yeo and Ong, 2023). For instance, being fired from a job, if judged to be due to one's own controllable mistakes, could result in regret; if evaluated to be unfair and due to someone else's intentional actions, would make one feel angry; and if appraised to be leaving a toxic work environment, could instead result in relief and even happiness. **The different dimensions along which people subjectively inter-**

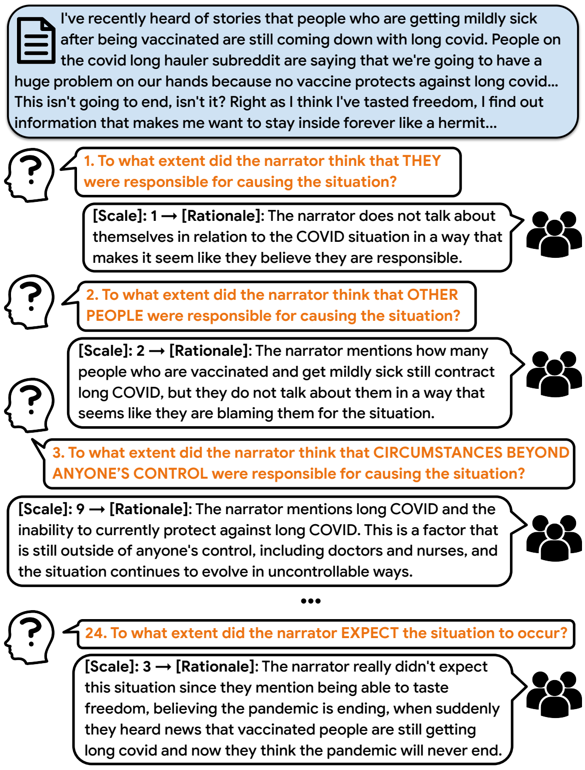


Figure 1: An example from COVIDET-APPRAISALS. The fact that the narrator is blaming nobody but circumstances beyond anyone's control for causing long-COVID contributes to their feeling of *sadness*. We showcase an annotation together with LLMs' responses in Appendix §A.

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 14418–14446
December 6-10, 2023 ©2023 Association for Computational Linguistics

pret or *appraise* the situation characterizes the specific emotions they feel (Moors et al., 2013).

Although emotion *detection* is a well-established NLP task (Strapparava and Mihalcea, 2007; Mihalcea and Strapparava, 2012; Wang et al., 2012; Lei et al., 2014; Abdul-Mageed and Ungar, 2017; Khanpour and Caragea, 2018; Liu et al., 2019; Sosea and Caragea, 2020; Demszky et al., 2020; Desai et al., 2020; Sosea et al., 2022), it mostly involves classification from text to emotion labels directly, skipping the appraisal step that is necessary to interpret why the emotion is experienced by an individual in a particular event. Hence, we do not yet have a data-driven understanding of these cognitive appraisals in textual data. Yet recent work has started to show its necessity: Hofmann et al. (2020) showed that appraisals are informative for an emotion detection model; Zhan et al. (2022) further recognized appraisals to be an integral part of emotion triggers, though appraisals were not explicit in their work.

This work aims at construing an empirical, explicit understanding of *perceived* cognitive appraisals in human readers and large language models (LLMs) alike, via a comprehensive 24 dimensions, along with their corresponding natural language rationales. A language model's capability of assessing cognitive appraisals reflects a more nuanced understanding of emotions, where it could contextualize individual subjectivity in responses to the same situation, while offering explanations ("they are feeling [*emotion*] because of [*appraisal*]"). This could be groundwork for emotional support agents, e.g., one capable of positive reframing (Ziems et al., 2022) or producing empathetic responses.

We first introduce COVIDET-APPRAISALS, a dataset of 24 appraisal dimensions annotated across 241 Reddit posts sourced from Zhan et al. (2022) about COVID-19. Each post was manually annotated with 24 appraisal dimensions from a recent meta-analysis covering all appraisal dimensions proposed and studied in the literature (Yeo and Ong, 2023). For each appraisal dimension, annotators not only rated the extent to which they perceived the narrator is experiencing the said dimension, but also provided a *rationale* in their own language to justify their rating selection. An example from COVIDET-APPRAISALS is shown in Figure 1.

COVIDET-APPRAISALS serves as an ideal testbed to evaluate the capability of a model to uncover implicit information for emotion understanding. Benchmarking on COVIDET-APPRAISALS, we evaluate the performance of LLMs to (1) provide Likert-scale ratings for the appraisal dimensions; and (2) generate natural language rationales for their ratings. The elicitation of the rationales can be seen as a way of probing (Le Scao and Rush, 2021; Gu et al., 2022), where we prefix a question with an elaborated situation. We evaluate a range of LLMs, including ChatGPT, Flan-T5 (Chung et al., 2022), Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023). With an extensive human evaluation of the natural language rationales from LLMs as well as our annotators, we find that ChatGPT performs on par with (and in some cases better than) human-annotated data; this opens a new avenue of investigation to improve its performance on emotion-related tasks (Kocoń et al., 2023). In comparison, other open-sourced LLMs fall short on this task, presenting a new challenge in the future development of emotionally intelligent open models.

We publicly release our annotated dataset COVIDET-APPRAISALS, *model outputs, and our human evaluation data at* `https://github.com/honglizhan/CovidET-Appraisals-Public`.

2 Background and Related Work

Cognitive Appraisal Theories. The cognitive appraisal theories of emotion state that emotions arise from an individual's subjective understanding and interpretation of situations that hold personal importance for their overall well-being (Arnold, 1960; Lazarus, 1966; Lazarus et al., 1980; Roseman, 1984; Scherer et al., 1984; Smith and Ellsworth, 1985; Weiner, 1985; Clore and Ortony, 2000; Roseman and Smith, 2001; Scherer et al., 2001; Sander et al., 2005; Ortony et al., 2022). In practical terms, people interpret and appraise situations along a range of different dimensions, and it is the specific manner in which they appraise their situations that give rise to the distinct emotions they experience. The primary focus of cognitive appraisal theories of emotions revolves around the identification of these appraisal dimensions that are associated with specific emotional experiences and how these dimensions contribute to distinguishing between different emotional states (Lazarus, 1993; Roseman, 1996; Scherer et al., 2001; Moors, 2010; Scherer and Moors, 2019).

While appraisal theorists agree on the importance of motivationally-relevant appraisals in triggering emotions, they have not reached a consensus on the specific appraisal dimensions that play a significant role in this process (Yeo and Ong, 2023). Various theories have put forth distinct sets of appraisal dimensions that are considered crucial in triggering and distinguishing emotions. From prior literature, Yeo and Ong (2023) identified and assembled a taxonomy of all appraisal dimensions that have been studied, and produced a condensed list of 24 cognitive appraisal dimensions which we focus on in this paper.

Cognitive Appraisals in NLP. Appraisals provide the necessary computational structure allowing for the distillation of real-life situations that depend on a multitude of factors into a (large but) finite set of appraisal dimensions (Ong et al., 2015). Despite its importance, however, few works have explored the implications of cognitive appraisals on emotions in NLP. Hofmann et al. (2020) experimented with a small set of cognitive appraisal dimensions (including *attention*, *certainty*, *effort*, *pleasantness*, *responsibility*, *control*, and *circumstance*) to assist the automatic detection of emotions in text, and found that accurate predictions of appraisal dimensions boost emotion classification performance. They introduced a dataset of $1,001$ sentences following the template "I feel [*emotion*], when ..." (average sentence length: 27 tokens). In comparison, our work covers a much wider range of 24 appraisal dimensions found in prior literature, over lengthy (176 tokens on average) Reddit posts that were natural and emotionally charged. We also collect natural language rationales as a key contribution to reveal human's in-depth understanding of such cognitive appraisals in context.

Recent studies (Zhan et al., 2022; Sosea et al., 2023) acknowledged both *what happened and how one appraised the situation* as inherent components of emotion triggers, although the appraisal of events was not explicit in their work. Instead we provide datasets and perform evaluation on appraisals explicitly, such that language models can build on this work to achieve a comprehensive and explicit understanding of cognitive appraisals from written text.

LLMs on Emotion-Related Tasks. Autoregressive LLMs have been explored extensively in emotion-related tasks such as sentiment analysis (Zhong et al., 2023; Qin et al., 2023; Susnjak, 2023), emotion recognition (Kocoń et al., 2023), disclosing the representation of human emotions encapsulated in LLMs (Li et al., 2023), and interpreting mental health analysis (Yang et al., 2023). However, few have tapped into the understanding of cognitive appraisals of emotions innate in LLMs. In this work, we dive into the extent to which LLMs comprehend the profound cognitive appraisals underlying emotions in situations, and further elicit natural language rationales from the language models to disclose the reason behind such predictions from the otherwise baffling black-box LLMs (Gilpin et al., 2018). Aligning with Marasović et al. (2020) who performed human evaluation on rationales generated by GPT, we additionally perform an in-depth human evaluation of the rationales from human annotators and LLMs alike on the novel task of providing natural language explanations for cognitive appraisals of situations that underlie narrators' emotional experiences.

3 The CovidET-Appraisals Dataset

CovidET-Appraisals contains 241 Reddit posts sampled from the CovidET dataset (Zhan et al., 2022), where the Reddit posts are sourced from `r/COVID19_support`. Each post is manually annotated with one or more of the 7 emotions: *anger*, *anticipation*, *joy*, *trust*, *fear*, *sadness*, and *disgust*. The 241 posts in CovidET-Appraisals have an average of 175.82 tokens and 2.67 emotions per post. From Yeo and Ong (2023)'s work, we identify 24 cognitive emotion appraisal dimensions (Table 1). We provide the instructions given to the annotators (including the full questions for each of these 24 dimensions) in Appendix §B.

Annotators. We recruited 2 linguistics students at a university to work on our annotation task; both of them are native speakers of English. Both annotators underwent training using a set of posts already annotated by our group. Throughout the annotation, we monitored the inter-annotator agreement and provided feedback on their work.

Instructions. Given a Reddit post from CovidET, annotators are asked to judge 24 emotion appraisal dimensions pertaining to how the narrator feels about and views the situation that they are going through (e.g., whether the narrator feels the situation they are in is something they could control). For each appraisal dimension, annotators need to select a Likert rating on the

ID	Abbrv.	Reader-Friendly Labels
1	*srsp*	*Self-responsibility*
2	*orsp*	*Other-responsibility*
3	*crsp*	*Circumstances-responsibility*
4	*pfc*	*Problem-focused coping*
5	*grlv*	*Goal Relevance*
6	*attn*	*Attentional activity*
7	*efc*	*Emotion-focused coping*
8	*scrl*	*Self-Controllable*
9	*ocrl*	*Other-Controllable*
10	*ccrl*	*Circumstances-Controllable*
11	*prd*	*Predictability*
12	*thr*	*Threat*
13	*pls*	*Pleasantness*
14	*crt*	*Certainty*
15	*gcnd*	*Goal Conduciveness*
16	*fair*	*Fairness*
17	*fex*	*Future expectancy*
18	*csn*	*Consistency with social norms*
19	*loss*	*Loss*
20	*fml*	*Familiarity*
21	*eff*	*Effort*
22	*chl*	*Challenge*
23	*civ*	*Consistency with internal values*
24	*exp*	*Expectedness*

Table 1: The 24 appraisal dimensions and their abbreviations we used throughout this paper. See Appendix §B for full questions for each dimension, and Figure 1 for an example of how the items for 1: *self-responsibility*, 2: *other-responsibility*, 3: *circumstances-responsibility*, and 24: *expectedness* were framed.

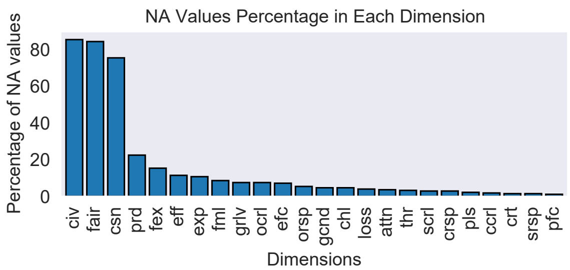


Figure 2: Percentage of *"not mentioned"* labels in each dimension in COVIDET-APPRAISALS.

scales of 1 to 9. A "*not mentioned*" (NA) option is provided in case the dimension being asked is absent in the given post. In addition, we also ask the annotators to provide rationales for their ratings in the form of *natural language explanations*.

On average, our trained annotators spent around 30 minutes to complete the annotation of one post. Owing to the immense effort involved, we doubly annotate 40 posts to measure inter-annotator agreement while leaving the rest annotated by one annotator.

Post-Processing and Aggregation. Given a fixed topic (COVID-19 in our case), it is highly likely that certain dimensions frequently don't apply (Yeo and Ong, 2023). This can be seen in Figure 2 which plots the percentage of NA labels: dimensions such as *civ* (consistency with internal values), *fair* (fairness), and *csn* (consistency with social norms) contain mostly NA labels (around 80%). Therefore, we remove these dimensions from subsequent analyses and evaluations of the dataset. **This results in a total of 21 *applicable* appraisal dimensions in COVIDET-APPRAISALS.**

We collected 241 posts in total. For the subset of 40 posts that are doubly annotated, we aggregate the Likert-scale ratings by taking the mean of each post's ratings for each appraisal dimension (if an annotator labels a dimension as NA, we then exclude the particular dimension of that post that they annotate). In terms of the rationales, we consider both rationales as ground truth references and use multi-reference metrics in our experiments.

Inter-Annotator Agreement. We report inter-annotator agreement on the Likert-scale ratings. Since there is no reliable, automatic way to evaluate natural language rationales (as discussed in §4), we evaluate them with human validation in §7.2.

To measure the agreement for selecting the NA label, we average the Fleiss' Kappa values (Fleiss, 1971; Randolph, 2005) across *all* 24 appraisal dimensions, yielding a value of 0.769 indicating substantial agreement (Artstein and Poesio, 2008).

For the 1-9 Likert-scale ratings, we report on the 21 applicable dimensions: (1) Spearman's ρ between our two annotators, calculated per dimension then averaged across all dimensions; (2) Krippendorff's alpha (using interval distance) (Krippendorff, 1980); and (3) mean absolute difference (*abs. delta*). Here the agreement is calculated if neither annotator gave a NA judgment. Krippendorff's alpha yields a value of 0.647 indicating substantial agreement (Artstein and Poesio, 2008). The average Spearman's correlation is 0.497 with significance, and the absolute delta values also have a small mean of 1.734. These measures indicate that while the task is subjective, annotators do align with each other with only a small difference compared to the scale of ratings (1-9). Agreement values differ by dimension, which we showcase in Appendix C.

4 Dataset Analysis

How do the scales distribute across dimensions and emotions? The distribution of the Likert-scale ratings is shown in Figure 3. The rat-

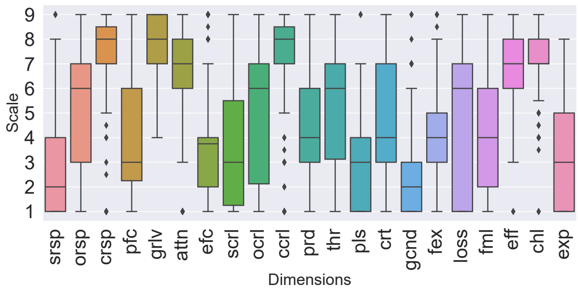


Figure 3: Distribution of the ratings for each dimension.

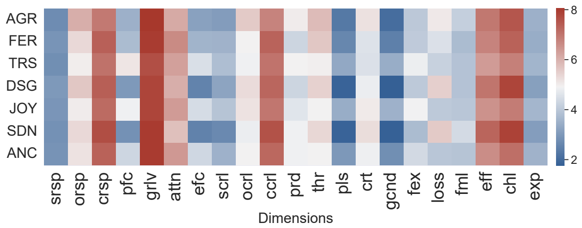


Figure 4: Mean Likert-scale ratings for each dimension in each emotion.

ings for some dimensions are consistent (e.g., dimensions *crsp* (circumstances-responsibility), *ccrl* (circumstances-controllable), and *chl* (challenge)), whereas for some other dimensions, the ratings have higher variance (e.g., dimensions *ocrl* (other-controllable) and *loss*).

We analyze the connections between our Likert-scale annotations and CovidET's emotion annotations. Figure 4 shows the mean Likert-scale rating for each dimension within each post with respect to the perceived emotion. While it is evident that most dimensions show consistency (the posts are all related to COVID-19), some emotions stand out distinctly in particular dimensions. For example, *trust* and *joy* have higher Likert-scale ratings on dimensions *pfc* (problem-focused coping) and *gcnd* (goal conduciveness) compared to other emotions, suggesting the inter-correlation between these appraisal dimensions with positive emotions. We further explore whether appraisal dimensions alone are indicative of perceived emotions already annotated in CovidET in Appendix §D.1.

What are the characteristics of the natural language rationales? On average, each rationale is 1.2 sentences (std.dev = 0.4) and 28.9 tokens (std.dev = 10.0) long. Following Marfurt and Henderson (2021), we also measure the abstractiveness of the rationales from our human annotators by calculating the percentage of novel bigrams in the rationales with respect to the Reddit posts and instructions (i.e., evaluating a specific appraisal dimension) that the annotators were given. As shown in Table 4, our human annotators attain a % of novel bigrams of 86.7%, indicating a high abstractiveness. We showcase the most prominent topics extracted from the annotated rationales using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in Appendix §D.2.

	Rationale		
	BLEU-4	ROUGE-L	BERTSc
Annotators	0.042	0.253	**0.357**
Baseline-P	**0.060**	**0.261**	0.336
Baseline-D	0.059	0.247	0.332

Table 2: Automatic measures of similarity on the natural language rationales of CovidET-Appraisals. Baseline-P denotes "baseline (same dimension, *different posts*)", and Baseline-D denotes "baseline (same post, *different dimensions*)".

Are rationales repetitive? We also look into automatic measures of similarity to assess how much rationales from different annotators, or from different dimensions/posts, differ from one another. Specifically, we calculate BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and re-scaled BERTScore (Zhang et al., 2019) between our two annotators' rationales. We establish 2 random baselines for comparison: (1) rationales of the same dimension from different posts; (2) rationales from different dimensions within the same post. In each case we report similarity between 3 randomly sampled rationales and the annotated ones.

Table 2 shows that the textual similarity in all conditions are somewhat low; the BLEU and ROUGE scores show that there is very little lexical overlap, although BERTScore shows higher semantic similarity between two annotators for the same dimension within the same post. Upon closer inspection, we observe that these commonly used automatic measures do not adequately capture semantic similarity in our dataset (see Appendix §D.3 for an example). This adds to the challenge of evaluating rationales; as a result, we resort to the human evaluation in §7.2.

5 Can LLMs understand emotional appraisals?

CovidET-Appraisals provides an ideal testbed that evaluates models' performance on predicting both the Likert ratings, as well as their natural language explanations. Using CovidET-Appraisals, we evaluate the zero-shot performance of LLMs in an attempt to evaluate their innate ability to comprehend emotional appraisals from social media text without in-context learning.

Models. We evaluate the following instruction-tuned LLMs[1]: **1) ChatGPT**, i.e., GPT-3.5-Turbo; **2) FLAN-T5-XXL (11B)** (Chung et al., 2022), which is the instruction fine-tuned version of T5 (Raffel et al., 2020); **3) Alpaca (7B, 13B)** (Taori et al., 2023) is fine-tuned from LLaMA (7B and 13B) (Touvron et al., 2023) on 52K instruction-following examples created with GPT text-davinci-003 in the manner of self-instruct (Wang et al., 2022); **4) Dolly-V2 (7B, 12B)** (Conover et al., 2023) is an instruction-tuned LLM trained on ~15k demonstrations consisting of both instructions and responses.

Prompts and Setup. The templates for prompting the LLMs are shown in Appendix Figure 17. After extensive experimentation, we found that only ChatGPT is able to generate both a rating and a rationale with a single prompt; this type of "1-step" prompting leads to ill-formed responses for other models. Thus, for models other than ChatGPT, we instead use a pipeline or "2-step" prompting similar to the strategy used in Press et al. (2022): we first elicit the rating for the appraisal dimension, then conditioned on the response for the rating we further elicit the rationale for the selection.

We carry out all our experiments on 4 Nvidia A40 GPUs. We use the HuggingFace Transformers (Wolf et al., 2020) library for model inference. We set the temperature value of all models to 0.1.[2] To enable a fair comparison of models, we sample from the LLMs five times with different model initializations and report average values for both scales and rationales.

6 Evaluation: Likert-Scale Ratings

We report model performance for Likert-scale ratings on the 21 *applicable* dimensions using two standard regression metrics: Mean Absolute Error (MAE) and Spearman's correlation. We treat the selection of the NA labels as a binary classification task and report F1 measures across *all* 24 dimensions. For the 40 gold examples that were doubly annotated by human annotators, we consider a dimension as NA when both annotators select the label.

	Scale		NA
	MAE	Spearman's ρ	F1
ChatGPT	**1.694**	$\mathbf{0.388}^{\dagger\dagger}$	**0.918**
Flan-T5	3.266	$0.225^{\dagger}$	0.852
Alpaca-7B	2.353	0.081	**0.918**
Alpaca-13B	3.872	−0.035	0.602
Dolly-7B	2.812	−0.013	0.645
Dolly-12B	2.747	0.022	0.711

Table 3: Experiment results from LLMs. † indicates $p < 0.1$ for Spearman correlation, and †† indicates $p < 0.05$. In addition, we also provide the results of the F1 score on measuring the agreement between the models' ratings and the gold ratings for selecting the "*not mentioned*" label across *all* 24 dimensions.

Results. To evaluate the performance, we clean the responses elicited from the LLMs. Specifically, we use regular expressions to extract the first numeric value ranging from 1-9 from the scale responses[3]. The results of the models' performance are shown in Table 3. We showcase examples of the models' responses in Appendix §A. Additional analyses of the LLMs' responses are shown in Appendix §G.

For the NA labels (Table 3, right), ChatGPT and Alpaca-7B score the highest with an F1 of 0.918. In general, the average performance across the language models we evaluate is 0.774 for F1, indicating these models are performant at predicting whether a dimension applies.

For the Likert-rating predictions, results show that ChatGPT-3.5 consistently yields the highest performance compared to the other language models, with a significant Spearman's correlation of 0.388 and an MAE of 1.694. We note that FLAN-T5-XXL is the second best-performing model. Alpaca and Dolly perform poorly on our task, with negative correlations with the gold labels[4]. Inter-

[1]While we have also experimented with non-instruction-tuned LLMs (including GPT-3 davinci and LLaMA (7B and 13B), they largely fail to generate sensible outputs for this task. We showcase examples of responses from non-instruction-tuned models in Appendix §A. For these reasons, we do not include their results in this paper.

[2]We experimented with higher temperatures on a validation set consisting of 10 Reddit posts annotated by our group which are not included in CovidET-Appraisals, and the models yielded worse and more unstable performance.

[3]For example, one of Alpaca-7B's scale responses is *"The narrator thought that Circumstances Beyond Anyone's Control were responsible for causing the situation to a moderate extent (4 on a scale of 1-9).</s>"*. After cleaning, the response is formatted to *"4"*.

[4]As shown in Appendix Figure 9, the ratings generated by

	LENGTH # TOKENS	ABSTRACTIVENESS %NOVEL BIGRAMS	AUTO EVAL BLEU-4	ROUGE-L	BERTSC	HUMAN EVAL FAC	REL	JUS	USE
ANNOTATORS	**28.9**	**86.7%**	——			0.73	**0.88**	**0.95**	0.72
CHATGPT	58.0	81.8%	**0.044**	0.224	**0.347**	**0.84**	**0.88**	0.93	**0.85**
FLAN-T5	45.3	16.0%	0.008	0.066	0.053	0.40	0.29	0.24	0.13
ALPACA-7B	48.6	71.9%	0.040	**0.230**	0.297	0.55	0.82	0.82	0.51

Table 4: Experiment results from LLMs. Additional evaluations of *all* language models (including Alpaca-13B, Dolly-7B, and Dolly-12B) are provided in Table 11. A more comprehensive report of the automatic metrics BLEU-4, ROUGE-L, and BERTSCORE is provided in Table 9, Appendix §F.

estingly, we notice a drop in performance when the size of the model parameters increases for Alpaca. The results highlight the challenging nature of our task, and the gap between open-sourced LLMs vs. ChatGPT (Gudibande et al., 2023).

Additionally, we also measure the systems' performance on all 24 appraisal dimensions, including the 3 appraisal dimensions where the NA rates are around 80%. Results revealed marginal change in performance across all LLMs. For most LLMs the performance dropped as expected: measured with Spearman's ρ, ChatGPT-3.5 ($\downarrow$ 0.018), Alpaca-7B ($\downarrow$ 0.008), and Dolly-12B ($\downarrow$ 0.007). On the other hand, the performance of FLAN-T5 ($\uparrow$ 0.005), Alpaca-13B ($\uparrow$ 0.027), and Dolly-7B ($\uparrow$ 0.020) increased.

7 Evaluation: Rationales

As rationalizing emotional appraisals with natural language is a novel task, we perform both automatic (§7.1) and human evaluation (§7.2).

7.1 Automatic Evaluation

We use commonly used automatic reference-based metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019), comparing generated rationales vs. annotated ones (in a multi-reference fashion).

Results. Similar to the performance in selecting Likert-scale ratings, ChatGPT remains the best-performing language model in providing natural language rationales (Table 4). The values ChatGPT achieves are lower than, though comparable to, those between different rationales from our two annotators. Alpaca-7B also achieves comparable performance in these automatic measures, despite its relatively poor capability in terms of selecting Likert-scale ratings. We note that FLAN-T5 lags behind considerably compared to ChatGPT and Alpaca-7B. We provide the additional auto-evaluation statistics for other LLMs including Dolly-7B, Dolly-12B, and Alpaca-13B in Appendix Table 11.

How long and how abstractive are the rationales generated by LLMs? In addition, we also measure the length and abstractiveness of the rationales generated by LLMs. Following the setup in §4, we evaluate abstractiveness using % of novel bigrams, comparing LLMs' generated rationales against the Reddit posts as well as the prompts (i.e., evaluating a specific appraisal dimension) they were given. As shown in Table 4, rationales generated by LLMs are at least 1.5x longer than those provided by our annotators, with ChatGPT being the most verbose. The LLMs also provide rationales that are more extractive compared to our annotators, with FLAN-T5 being the most extractive.

7.2 Human Evaluation

Data. Because the natural language rationales are explanations for a particular rating, we only evaluate and analyze LLM-generated rationales when the model made a near-correct prediction of the Likert-scale rating for that particular dimension compared against the gold human ratings. Specifically, we sample the *intersection* of (post, dimension) tuples where the 3 *best-performing* LLMs' (i.e., ChatGPT, FLAN-T5, and Alpaca-7B) ratings fall in the range of an absolute difference of 1 to one of the annotated scale-ratings. In cases where there are 2 gold annotations for a particular dimension, both are evaluated. In Appendix §F we also show the human evaluation of rationales for such intersection of *all* LLMs. We additionally evaluate **human-written rationales** as well, and we mix those (in random order) with LLMs' responses.

The above desiderata results in an evaluation

the language models (specifically, Alpaca-7B and Dolly-12B) for some of the dimensions lack variance (i.e., they gave a constant rating for certain appraisal dimensions). Therefore, the Spearman correlation is set to zero in these dimensions, indicating no correlation.

of 108 rationales annotated by human annotators and 65 natural language rationales from each LLM. The evaluation covers 19 out of the 21 applicable dimensions (no such overlap is found for dimensions *crsp* (*circumstances-responsibility*) and *pls* (*pleasantness*)). Moreover, we make sure that there are no ground truth labels annotated by the human annotators in which the rating is NA.

Instructions. Given a Reddit post and the scale provided by the human annotators or the LLM (blinded to the annotators), annotators are asked to judge the rationales pertaining to the emotion appraisal dimension regarding the post as well as the stated scale. The rationales are distributed to annotators at random. We evaluate the natural language rationales based on the following criteria. In Appendix §H, We provide the detailed instructions and examples given to the annotators, together with the layout of the human evaluation task.

1) Factuality: For the rationale, the model may not generate something that is factual: sometimes it generates rationales for the sole purpose of justifying its answer (Ye and Durrett, 2022). Therefore, we include the aspect of *hallucination and factuality* as one of our evaluation criteria, and ask evaluators whether the rationale faithfully reflects what's stated in the post. Options of "*Yes*", "*Minor Error*", and "*No*" are provided.

2) Relevance: We evaluate whether the rationale directly addresses the specific appraisal dimension question that is being asked about the post. We ask evaluators on a Likert-scale of 1 to 5, with 1 being "*least relevant*" and 5 being "*most relevant*", whether the rationale focuses on the specific aspect of the post that is being appraised, and whether it strays off-topic or provides irrelevant information.

3) Justification: We ask human evaluators whether the rationale justifies the selected scale by adequately explaining why the selected rating scale is the most appropriate or relevant one to use for the aspect being evaluated. Annotators need to select either "*Yes*" or "*No*".

4) Usefulness: Finally, we evaluate whether the rationale provides useful or informative insights or explanations of useful information pertaining to the appraisal dimension being judged. Options of "*Yes*", "*Maybe*", and "*No*" can be selected.

Annotators. We recruit annotators from the Amazon Mechanical Turk (MTurk) to work on our human evaluation task. The crowd workers were involved in a pre-annotation *qualification as well as training* process before commencing the evaluation of the natural language rationales. We assign 2 crowd workers per natural language rationale evaluation. We ensure that the crowd workers earn a minimum salary of $10 per hour.

We report the inter-evaluator agreement using Krippendorff's Alpha with interval distance in Table 5, showing substantial agreement (Artstein and Poesio, 2008) across all criteria.

Label Transformation. For the convenience of measuring inter-annotator agreement as well as interpreting the results, we convert the labels of each criterion to numeric values within the range of 0 to 1. Specifically, for criteria *Factuality*, *Justification*, and *Usefulness*, "*Yes*" is converted to 1, "*Minor Error/Maybe*" to 0.5, and "*No*" to 0. As for the criterion *Relevance* which is judged on a 5-scale Likert rating, we map the Likert scale of 1 into 0, 2 into 0.25, 3 into 0.5, 4 into 0.75, and 5 into 1.

Results. The result of the mean ratings for each criterion from the human evaluation task is provided in Table 4. We provide box plots of the ratings as well as the human evaluation results for the rationales from all 6 LLMs in Appendix §F.

From Table 4 we observe that our human annotators and ChatGPT provide natural language rationales of the highest quality among all models, according to human evaluators. Surprisingly, we find ChatGPT performs on par with our human annotators, with (slightly) better performance in terms of *factuality* and *usefulness*. This can be attributed to the verbosity and extractiveness of ChatGPT (as shown in Table 4), especially in dimensions where the scale rating is low. We showcase an example in Appendix §I.

Alpaca-7B attains lower results compared to the other LLMs, especially in terms of the criteria *factuality* and *usefulness*. FLAN-T5, on the other hand, ranks the worst on all criteria among the LLMs. Further analysis reveals that FLAN-T5 occasionally generates responses for natural language rationales that are the same as its scale answers, resulting in irrelevant and useless rationales.

8 Conclusion

To achieve a more accurate and holistic understanding of emotions from written text, NLP models need to work towards understanding the subjective cognitive appraisals of emotions underlying

	FAC	REL	JUS	USE
EVALUATORS	0.590	0.718	0.576	0.668

Table 5: Inter-annotator agreement statistics for the human evaluation task, measured using Krippendorff's Alpha with interval distance.

situations. In this work, we construe an empirical and explicit understanding of *perceived* cognitive appraisals in human readers and LLMs alike. We present COVIDET-APPRAISALS, a dataset of 241 Reddit posts annotated with a comprehensive range of 24 subjective cognitive appraisals that follow a situation, along with their corresponding natural language rationales. Experiments reveal that COVIDET-APPRAISALS is a vital resource to evaluate the capability of a language model to uncover implicit information for emotional understanding. Our thorough evaluation of LLMs' performance on assessing emotion appraisal dimensions emphasizes that COVIDET-APPRAISALS is a challenging benchmark, and our in-depth human evaluation of the natural language rationales indicates potential areas of improvement (e.g., improving the *factuality* and *usefulness* of the rationales) for open-source LLMs.

Limitations

This work presents a new dataset entitled COVIDET-APPRAISALS to evaluate LLMs' capability in cognitive emotion appraisals. Due to the highly demanding nature of our task (e.g., the same situation can result in different subjective evaluations), COVIDET-APPRAISALS is annotated by 2 annotators. Future work can explore a larger pool of annotators. Furthermore, it should be acknowledged that COVIDET-APPRAISALS is restricted to social media posts during the COVID-19 pandemic, and they are written in English solely. This makes it challenging to evaluate LLMs' ability in other domains as well as languages. Also, we note the appraisals we collect are from the *perceived* end, which are not subjective appraisals from the narrators and authors themselves.

We note that the size of COVIDET-APPRAISALS is relatively small. We have not intended this dataset to be one for supervised model training but rather a very high-quality dataset for evaluation (since this is the first dataset of its kind). A key reason is that the collection of appraisal annotations is both challenging and time-consuming: we have 24 dimensions to analyze per post, and the annotation for one post for one trained annotator takes half an hour. Future work may establish the validity of training data obtained from LLMs, and explore approaches such as distillation.

In addition, we experiment with LLMs under a zero-shot setup only, while we highlight that this is the first work towards the assessment of cognitive appraisals of emotions in language models, and it lays the foundation for future research on deciphering the intrinsic emotional dynamics that remain unexplored in current state-of-the-art models. We believe that this warrants a careful construction of the dataset with thorough analysis; and we leave these interesting engineering questions to future work.

Acknowledgements

This research was partially supported by National Science Foundation (NSF) grant IIS-2107524. We thank Kathryn Kazanas and Keziah Kaylyn Reina for their dedication and hard work on the annotation of COVIDET-APPRAISALS. We also thank our reviewers for their insightful feedback and comments.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Magda B Arnold. 1960. *Emotion and personality.* Columbia University Press.

Anshika Arora, Pinaki Chakraborty, M. P. S. Bhatia, and Prabhat Mittal. 2021. Role of Emotion in Excessive Use of Twitter During COVID-19 Imposed Lockdown in India. *Journal of Technology in Behavioral Science*, 6(2):370–377.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, 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. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Gerald L Clore and Andrew Ortony. 2000. Cognition in emotion: Always, sometimes, or never. *Cognitive neuroscience of emotion*, pages 24–61.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Phoebe C Ellsworth and Klaus R Scherer. 2003. *Appraisal processes in emotion.* Oxford University Press.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. DREAM: Improving situational QA by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

Klaus Krippendorff. 1980. *Content analysis : an introduction to its methodology*. Sage commtext series. Sage Publications, Beverly Hills.

Richard S Lazarus. 1966. *Psychological stress and the coping process.* McGraw-Hill.

Richard S Lazarus. 1993. From psychological stress to the emotions: A history of changing outlooks. *Annual review of psychology*, 44(1):1–22.

Richard S Lazarus, Allen D Kanner, and Susan Folkman. 1980. Emotions: A cognitive–phenomenological analysis. In *Theories of emotion*, pages 189–217. Elsevier.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, 37:438–448.

Ming Li, Yusheng Su, Hsiu-Yuan Huang, Jiali Cheng, Xin Hu, Xinmiao Zhang, Huadong Wang, Yujia Qin, Xiaozhi Wang, Zhiyuan Liu, and Dan Zhang. 2023. Human emotion knowledge representation emerges in large language models and supports discrete emotion inference. *arXiv preprint arXiv:2302.09582*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6293–6298, Hong Kong, China. Association for Computational Linguistics.

Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2810–2829, Online. Association for Computational Linguistics.

Andreas Marfurt and James Henderson. 2021. Sentence-level planning for especially abstractive summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 1–14, Online and in Dominican Republic. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea. Association for Computational Linguistics.

Agnes Moors. 2010. *Theories of emotion causation: A review*. Psychology Press.

Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124.

Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne C. Reddan, Isabella Kahhalé, Alison Mattek, and Jamil Zaki. 2019. Modeling emotion in complex stories: The stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12:579–594.

Desmond C. Ong, Jamil Zaki, and Noah D. Goodman. 2015. Affective cognition: Exploring lay theories of emotion. *Cognition*, 143:141–162.

Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

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

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Ira J Roseman. 1984. Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology*.

Ira J Roseman. 1996. Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition & Emotion*, 10(3):241–278.

Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pages 3–19.

David Sander, Didier Grandjean, and Klaus R Scherer. 2005. A systems approach to appraisal mechanisms in emotion. *Neural networks*, 18(4):317–352.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44:695 – 729.

Klaus R Scherer and Agnes Moors. 2019. The emotion process: Event appraisal and component differentiation. *Annual review of psychology*, 70:719–745.

Klaus R Scherer et al. 1984. On the nature and function of emotion: A component process approach. *Approaches to emotion*, 2293(317):31.

KR Scherer, A Schorr, and T Johnstone. 2001. Appraisal theory: Overview, assumptions, varieties.

Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Tiberiu Sosea, Chau Pham, Alexander Tekle, Cornelia Caragea, and Junyi Jessy Li. 2022. Emotion analysis and detection during COVID-19. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6938–6947, Marseille, France. European Language Resources Association.

Tiberiu Sosea, Hongli Zhan, Junyi Jessy Li, and Cornelia Caragea. 2023. Unsupervised extractive summarization of emotion triggers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9550–9569, Toronto, Canada. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Teo Susnjak. 2023. Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. *arXiv preprint arXiv:2302.06474*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and A. Sheth. 2012. Harnessing twitter “big data” for automatic emotion identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4):548.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Gerard Yeo and Desmond C. Ong. 2023. A meta-analytic review of the associations between cognitive appraisals and emotions in cognitive appraisal theory. *PsyArXiv*.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022. Why do you feel this way? summarizing triggers of emotions in social media posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.

A Dataset Example and LLM Responses

In Figure 10, Figure 11, and Figure 12, we showcase an annotation from COVIDET-APPRAISALS together with LLMs' responses regarding dimension 3 *crsp* (circumstances-responsibility). In addition to LLMs evaluated in this paper (including ChatGPT, FLAN-T5-XXL, Alpaca (7B, 13B), and Dolly-V2 (7B, 12B)), we also present responses elicited from other non-instruction-tuned models such as GPT-3-davinci (a vanilla base model of GPT-3) and LLaMA (7B, 13B) (Touvron et al., 2023) using the "2-step" prompting template given in Figure 17. As the example shows, these non-instruction-tuned LLMs perform poorly on our task of cognitive emotion appraisal, generating nonsensical responses for both selecting Likert-scale ratings as well as providing natural language rationales.

B Dataset Annotation Framework

We provide the instructions given to the annotators in Figure 13. In addition, we also provide the layout for the annotation task (which includes the full questions for each of the 24 cognitive emotion appraisal dimensions abbreviated in Table 1) in Figures 14, 15, 16.

C Inter-Annotator Agreement by Dimension in COVIDET-APPRAISALS

To better understand the inter-annotator agreement pertaining to each emotion appraisal dimension in COVIDET-APPRAISALS, we measure Spearman's ρ and Krippendorff's alpha on each of the 21 applicable dimensions. We provide the inter-annotator agreement statistics per dimension in Figure 5. As the plot shows, the human annotators have strong agreement on dimensions such as *efc* (emotion-focused coping) and *pfc* (problem-focused coping), whilst disagreeing with each other most often on dimensions *grlv* (goal relevance), *exp* (expectedness), and *loss*. This can be attributed to the nature of our domain: in these Reddit posts, the narrator is mainly sharing their experiences in life around COVID-19, while preserving doubts about the future.

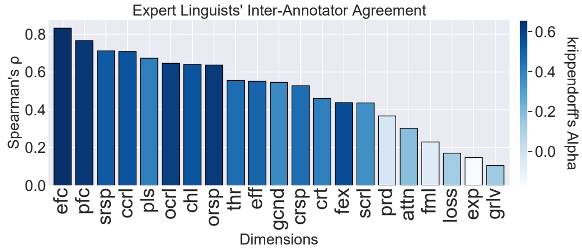


Figure 5: Inter-annotator agreement of the Likert-scale ratings within each dimension. The dimensions are ranked by the order of Spearman's ρ, and the colors indicate the inter-annotator agreement measured by Krippendorff's alpha using interval distance.

	AGR	DSG	FER	JOY	SDN	TRS	ANC	AVG
F1	0.18	0.13	0.40	0.26	0.29	0.06	0.23	**0.22**

Table 6: F1 scores of each emotion using the trained logistic regression model on the test set.

D Additional Dataset Analyses

D.1 Are the Dimensions Informative for Emotions?

The cognitive appraisal theories provide insights into the nature of the appraisal dimensions in distinguishing various emotions (Hofmann et al., 2020; Yeo and Ong, 2023): while different individuals may appraise the same situation distinctively, they are more likely to experience the same emotion when a consistent appraisal pattern emerges. For example, the cognitive dimension *pls* (pleasantness) is often linked to joy, but unlikely to be associated with disgust (Smith and Ellsworth, 1985). Therefore, specific emotions are hypothesized to stem from corresponding appraisal patterns (Yeo and Ong, 2023). By understanding how individuals appraise the situations they experience, we can subsequently make predictions regarding their emotional state. As a result, appraisal dimensions are valuable in differentiating emotional states, especially in cases where the emotions are highly interchangeable (e.g., *disgust* and *anger*).

Here, using the cognitive appraisal dimensions annotated in COVIDET-APPRAISALS, we further explore and validate whether these appraisal dimensions alone are indicative of perceived emotions already annotated in COVIDET. While in the ideal scenario, both the appraisal and the objective event need to be present for emotion prediction, this small experiment will allow us to gauge which dimensions are more likely discriminative for a particular emotion. For each of the 7 emo-

ID	Abbrv.	Reader-Friendly Labels	Anger	Fear	Joy	Sadness	Disgust
1	*srsp*	Self-responsibility		+	+	+	
2	*orsp*	Other-responsibility	+			+	+
3	*crsp*	Circumstances-responsibility		+		+	
4	*pfc*	Problem-focused coping	-	-	$+^{\dagger\dagger}$		
5	*grlv*	Goal Relevance	$+^{\dagger}$	+		+	+
6	*attn*	Attentional activity		+	+	+	+
7	*efc*	Emotion-focused coping		-	+	-	
8	*scrl*	Self-Controllable		-	+	-	
9	*ocrl*	Other-Controllable					+
10	*ccrl*	Circumstances-Controllable		+		+	
11	*prd*	Predictability	-	-		-	
12	*thr*	Threat	$+^{\dagger}$	+	-	+	+
13	*pls*	Pleasantness	-	-	+	-	-
14	*crt*	Certainty		-	+	-	
15	*gcnd*	Goal Conduciveness	-		+	-	+
17	*fex*	Future expectancy			+		
19	*loss*	Loss	+	+	-	+	
20	*fml*	Familiarity		-		-	
21	*eff*	Effort		+	-	+	
22	*chl*	Challenge					
24	*exp*	Expectedness					+

Table 7: Cognitive emotion appraisal dimensions that are predictive of emotions (including *anger*, *fear*, *joy*, *sadness*, and *disgust*), identified by a recent meta-analysis conducted by Yeo and Ong (2023). $+$ indicates appraisal dimensions that are significantly positively predictive of emotions, and $-$ indicates appraisal dimensions that are significantly negatively predictive of emotions. We highlight in red the indicative appraisal dimensions captured by our logistic regression models that are in line with Yeo and Ong (2023)'s findings. † signifies weights in our logistic regression models with $p < 0.1$, and †† signifies significant weights with $p < 0.05$.

tion classes labeled in COVIDET, we train a logistic regression model using the scales of the annotated 21 applicable appraisal dimensions as features. We split COVIDET-APPRAISALS using a random 80:20 train-test partitioning, and aggregate the Likert-scale ratings for the 40 posts that are doubly annotated by our human annotators following the aggregation setup discussed in §3. We downsample the training data for each logistic regression model to handle class imbalance issues. In addition, we encode the *"not mentioned"* (NA) labels as an independent real-valued feature, and substitute their values with 0. To prevent features of different scales or magnitudes from having a disproportionate influence on the models, we Z-normalize the scale ratings within each dimension for each annotator.

The F1 scores for each emotion using the trained logistic regression models on the test set are reported in Table 6. We observe that the models are most capable at predicting emotions such as *fear* and *sadness*, whilst performing poorly on emotions *disgust* and *trust*. This is possibly due to the domain of our dataset: in COVIDET, *fear* and *sadness* are the most commonly found emotions whereas *disgust* and *trust* are scarcely present. On average, the classifiers achieve an average F1 of 0.22 on the test set across all emotions.

To reveal the appraisal dimensions that are indicative of each emotion, we examine the weights from the trained logistic regression models. Specifically, we aim to validate the emotion appraisal dimensions that Yeo and Ong (2023) identified to be predictive of emotions (including *anger*, *fear*, *joy*, *sadness*, and *disgust*) from prior studies in psychology. In Table 7, we show the appraisal dimensions found to be either positively predictive $(+)$ or negatively predictive $(-)$ of emotions. Please note that these indications are extracted from a recent meta-analysis from Yeo and Ong (2023) with significance ($p < 0.05$). In Table 7, we highlight the indicative appraisal dimensions captured by our logistic regression models that are in line with Yeo and Ong (2023)'s findings. We observe a certain degree of overlap between Yeo and Ong (2023)'s identified emotion appraisal dimensions that are predictive of emotions and those captured by our logistic regression models. It should be noted that some appraisal dimensions may not be useful for all emotions included in Table 7, since in COVIDET there are no Reddit posts annotated with neutral emotions: for example, as shown in Table 7, *crsp* (circumstances-responsibility) is found to be positively indicative for *fear* and *sadness*, while neutral

srsp	*orsp*	*crsp*	*pfc*	*grlv*	*attn*	*efc*
believe	responsible	control	cope	finds	attend	cope
responsible	people	believes	believe	concerns	believes	emotionally
does	believes	circumstances	doesn	highly	need	somewhat
doesn	does	covid	coping	relevant	want	feeling
causing	covid	responsible	having	covid	believe	struggling
focused	vaccinated	blame	vaccine	infected	covid	believe
reaction	believe	delta	believes	stuck	advice	covid
believes	somewhat	outside	covid	dose	asking	believes
somewhat	blame	pandemic	difficult	ending	pandemic	doesn
vaccinated	causing	worried	time	pandemic	trying	coping
scrl	***ocrl***	***ccrl***	***prd***	***thr***	***pls***	***crt***
control	people	control	happen	threatened	finds	uncertain
believe	control	covid	believe	covid	unpleasant	unsure
does	believes	believes	predict	feels	feeling	certain
believes	wait	circumstances	doesn	does	covid	consequences
doesn	vaccine	outside	covid	express	pandemic	vaccine
covid	covid	delta	don	feeling	worried	covid
feel	somewhat	understands	unable	health	pleasant	understand
vaccine	does	understand	prediction	threat	confused	somewhat
vaccinated	believe	believe	makes	somewhat	feel	delta
pandemic	september	pandemic	information	sense	vaccine	fully
gcnd	***fex***	***loss***	***fml***	***eff***	***chl***	***exp***
want	worse	sense	subject	effort	finds	occur
finds	better	does	information	deal	challenging	did
inconsistent	believe	express	meaning	mental	covid	expect
covid	does	loss	advice	believes	vaccinated	mentions
highly	believes	lost	asking	lot	highly	somewhat
wants	getting	believes	mentions	exert	pandemic	expected
vaccinated	covid	covid	unfamiliar	try	vaccine	covid
don	delta	pandemic	familiar	believe	worried	expecting
feel	worried	vaccinated	covid	covid	delta	mention
trying	variant	opportunity	somewhat	need	variant	vaccinated

Table 8: LDA results on the annotated rationales for each appraisal dimension.

for all other emotions. However, when compared to neutral emotions (i.e., in texts where no emotions are present), *crsp* (circumstances-responsibility) may be a negative indicator for *disgust*. Therefore, experimenting with COVIDET-APPRAISALS may not reveal the extensive range of appraisal dimensions indicative of each emotion. Further investigations are needed to explore the predictability of these appraisal dimensions for emotions compared against neutral emotions.

D.2 Topic Variations in Rationales

We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to extract topics from the natural language rationales annotated in COVIDET-APPRAISALS. Stop-words such as common English function words and words that occur frequently in our instructions (e.g., *narrator*, *situation*) are removed prior to the topic modeling. The most prominent topic extracted by the LDA model for each dimension is shown in Table 8. We notice clear patterns of topics related to the appraisal dimension being assessed. For example, in dimension *crsp* (circumstances-responsibility) we observe narrators of Reddit posts worrying about and blaming Delta, a COVID-19 variant, for causing the status quo, whereas in dimension *fml* (familiarity) we note people are generally unfamiliar with the situation, as they are prone to seek advice and probe for information on the forum.

D.3 An Example of Semantic Similarity

As discussed in §4, commonly used automatic measures such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) do not adequately capture semantic similarity in COVIDET-APPRAISALS. Taking the post in Figure 1 for example. Both rationales for dimension 24, namely *"The narrator mentions how people who are vaccinated and mildly sick are still experiencing long COVID symptoms. They seem surprised by the continued COVID symptoms people are experiencing and how the situation seems to evolve."* and *"The narrator really didn't expect*

	BLEU			ROUGE			BERTSCORE	
	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	RE-SCALED
CHATGPT	**0.147**	**0.078**	**0.044**	**0.317**	**0.111**	0.224	**0.890**	**0.347**
ALPACA-7B	0.136	0.069	0.040	0.292	0.101	**0.230**	0.881	0.297
ALPACA-13B	0.007	0.004	0.003	0.019	0.005	0.017	0.842	0.066
DOLLY-7B	0.067	0.034	0.020	0.185	0.047	0.142	0.858	0.157
DOLLY-12B	0.086	0.043	0.024	0.223	0.066	0.165	0.865	0.199
FLAN-T5-XXL	0.026	0.014	0.008	0.091	0.018	0.066	0.840	0.053

Table 9: The full rationale statistics measured for LLMs' responses against the gold annotations, measured across 5 independent runs.

	FAC	REL	JUS	USE
EVALUATORS	0.721	0.711	0.632	0.672

Table 10: Inter-annotator agreement statistics for the human evaluation task, measured using Krippendorff's Alpha with interval distance.

this situation since they mention being able to taste freedom, believing the pandemic is ending, when suddenly they heard news that vaccinated people are still getting long covid and now they think the pandemic will never end." convey the reasons for why the narrator fails to expect the situation to occur. However, the automatic metrics reveal low agreement between these two rationales, with a BLEU-4 score of 0.018, ROUGE-L of 0.231, and a re-scaled BERTSCORE of 0.237. This finding is in line with work showing the challenges of *evaluating* generation (Gehrmann et al., 2021; Celikyilmaz et al., 2020); we similarly conclude that automatic evaluation metrics may poorly reflect the correctness of a rationale for a subjective emotion appraisal dimension.

E Prompt Templates

The templates for prompting the LLMs are shown in Figure 17. We use "1-step" prompting to elicit both a rating and a rationale with a single prompt from ChatGPT. For all other language models, we apply "2-step" prompting, which first elicits the rating for the appraisal dimension, then conditioned on the response for the rating we further elicit the rationale for the selection.

F Full LLM Rationale Measures

Rationale Automatic Evaluation. We provide the full statistics of the automatic rationale agreement measured using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) for the ***all*** 6 **LLMs' responses** against the gold annotations in Table 9.

As discussed in §7.1, ChatGPT is the most performant language model in providing natural language rationales, with values from these metrics comparable to those between different rationales from our two annotators. Alpaca-7B also achieves comparable performance in these automatic measures, despite its relatively poor capability in terms of selecting Likert-scale ratings.

In addition, we observe that other language models such as FLAN-T5 and Dolly lag behind considerably compared to ChatGPT and Alpaca-7B. Enchantingly, the automatic metrics suggest that Alpaca-13B is the worst language model among our LLMs under assessment, with a markable degradation from Alpaca-7B. Further investigation reveals that Alpaca-13B tends to respond with "*Tell us why.</s>*" when prompted to generate the natural language rationale for the Likert-scale rating it selects, which takes up more than 84% of its rationale responses. The debasement of the Alpaca model in spite of the increase in the model's scale raises questions regarding the scaling law in our current task of appraising cognitive emotion dimensions in context.

Rationale Human Evaluation. We provide the box plots of the results from the human evaluation for *the most-performant* 3 language models (i.e., ChatGPT, Alpaca-7B, and FLAN-T5) in Figure 6.

Furthermore, we also provide the results for the human evaluation regarding ***all*** 6 **LLMs** assessed in this paper. Following the setup in §7.2, we evaluate and analyze LLM-generated rationales when the model made a near-correct prediction of the Likert-scale rating for that particular dimension compared against the gold human ratings. Specifically, we sample the *intersection* of dimensions (post, dimension) tuples where *all* 6 LLMs' (i.e., ChatGPT, FLAN-T5, Alpaca-7B, Alpaca-13B, Dolly-7B, and Dolly-12B) ratings fall in the range of an abso-

	Length	Abstractiveness	Auto Eval			Human Eval			
	# Tokens	%Novel Bigrams	BLEU-4	ROUGE-L	BERTSc	FAC	REL	JUS	USE
Annotators	28.9	86.7%	——			0.68	**4.43**	**0.92**	0.77
ChatGPT	58.0	81.8%	**0.044**	0.224	**0.347**	**0.88**	4.42	0.85	**0.88**
Flan-T5	45.3	16.0%	0.008	0.066	0.053	0.44	2.27	0.25	0.19
Alpaca-7B	48.6	71.9%	0.040	**0.230**	0.297	0.57	4.23	0.79	0.64
Alpaca-13B	19.7	10.9%	0.003	0.017	0.066	0.03	1.13	0.02	0.02
Dolly-7B	79.7	51.3%	0.020	0.142	0.157	0.32	2.44	0.21	0.18
Dolly-12B	73.3	55.1%	0.024	0.165	0.199	0.38	2.79	0.56	0.38

Table 11: Experiment results from LLMs. We report the average performance across five independent runs. A more comprehensive report of the automatic metrics BLEU-4, ROUGE-L, and BERTScore is provided in Table 9, Appendix §F.

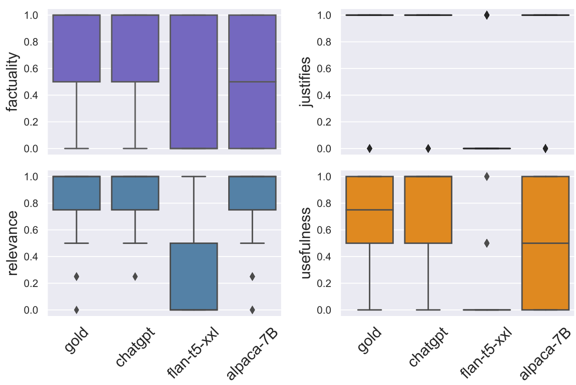


Figure 6: Box plots of the results from the human evaluation task for *the most-performant* 3 LLMs (i.e., ChatGPT, Alpaca-7B, and FLAN-T5).

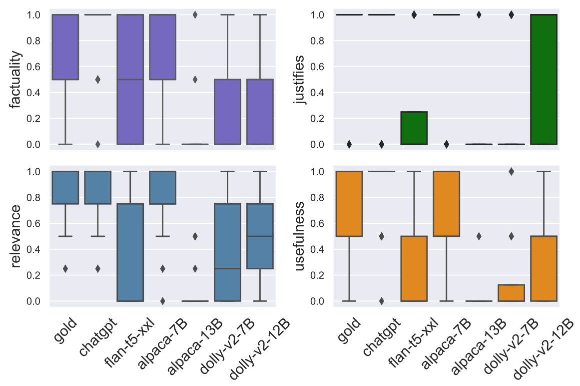


Figure 7: Box plots of the results from the human evaluation task for *all* 6 LLMs.

lute difference of 1 to *one of* the annotated scale-ratings. This results in 30 rationales annotated by human annotators and 26 natural language rationales from each LLM. We report the inter-evaluator agreement using Krippendorff's Alpha with interval distance in Table 10, which shows substantial agreement (Artstein and Poesio, 2008) across all criteria.

Results from the human evaluation for *all* 6 LLMs are reported in Table 11. We observe that apart from ChatGPT and Alpaca-7B, all other LLMs including FLAN-T5, Alpaca-13B, Dolly-7B, and Dolly-12B achieve similarly low performance on providing natural language rationales for cognitive emotion appraisals. We provide the box plots of the results from the human evaluation for *all* 6 language models in Figure 7.

G Model Responses Analyses

The LLMs' performance in terms of Likert-scale rating selections measured using Spearman correlation and Krippendorff's alpha against the gold annotations are shown in Figure 8. Additionally, the box plots for each LLM's Likert-scale ratings are shown in Figure 9.

H Human Evaluation Framework

We provide the instructions given to the human evaluators of the rationales (described in §7.2) in Figure 18 and Figure 19. Additionally, we showcase the human evaluation task layout in Figure 20.

I Why Does ChatGPT Perform (Slightly) Better Than Human Annotators in Providing Rationales?

As discussed in §7.2, ChatGPT was scored slightly higher in terms of *factuality* and *usefulness* on providing natural language rationales than our human annotators, according to human evaluators. This can be attributed to ChatGPT's wordiness and extractiveness (as shown in Table 4), especially in cognitive emotion appraisal dimensions where the scale rating is low. As an example, we showcase in Table 12 where both ChatGPT and our human annotator give the same rating for a dimension, but ChatGPT scores higher than our human experts on

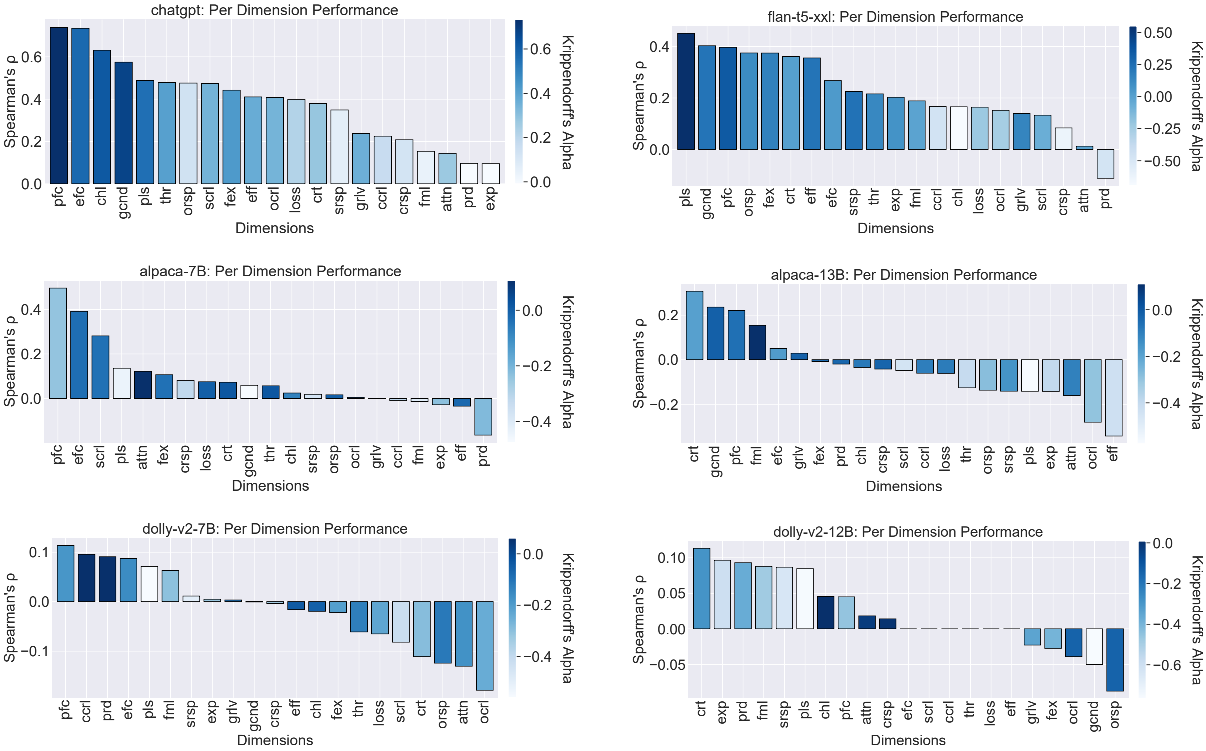


Figure 8: LLMs' performance in terms of Spearman correlation and Krippendorff's alpha (using interval distance) against the gold annotations within each group of dimensions (averaged performance across 5 independent runs).

metrics *factuality* and *usefulness*.

As shown in the example, given the same Reddit post as well as the instruction to evaluate the cognitive emotion appraisal dimension *orcl* (other-controllable), both our human annotator and ChatGPT give a Likert rating of 1 indicating a small extent to which the narrator believes other people were controlling what was happening in the situation. Similarly, in their natural language rationales, both our annotator and ChatGPT express that there is no mention of other people controlling the situation in the Reddit post. However, in the post, although to a small degree, the narrator is hinting that other people are in control of the situation since their school is deciding on the mask policy for the school year. This lowers the human annotation to a mean *factuality* rating of 0.25 and a mean *usefulness* rating of 0.50. On the other hand, ChatGPT iterates and summarizes the information provided in the post, mentioning that the narrator is *"seeking advice on where to find masks that are more suitable for their needs"*. The provision of such information thus leads to a higher score than our annotator, resulting in a mean rating of 0.75 for both *factuality* and *usefulness*.

REDDIT POST

I want to clarify that I am in no way shape an anti masker, I believe we do need them during the pandemic and any future outbreaks. However, the masks are a nightmare for me personally. They trigger off my migraines, I could hardly concentrate at school because of it. I also have a sensory disability, which makes wearing masks 10 times harder. Everything is brighter, louder, and everything feels more crowded when I wear them. Its horrible. And it looks like we have to wear them for the next school year, and its my final year. I cant risk it. So I want to know if any you know where I can get masks that are migraine/sensory friendly? While still doing the job a normal medical mask does. Thanks.

APPRAISAL DIMENSION *ocrl*

To what extent did the narrator think that OTHER PEOPLE were controlling what was happening in the situation?

HUMAN ANNOTATIONS

→ SCALE: 1

→ RATIONALE: *The narrator does not mention anyone else as controlling what's happening in the situation.*

FACTUALITY: 0.25 **(Minor Error + No)**

USEFULNESS: 0.50 **(Maybe + Maybe)**

CHATGPT'S RESPONSES

→ SCALE: 1

→ RATIONALE: *The text does not suggest that other people are controlling the situation. The narrator is simply seeking advice on where to find masks that are more suitable for their needs.*

FACTUALITY: 0.75 **(Yes + Minor Error)**

USEFULNESS: 0.75 **(Yes + Maybe)**

Table 12: An example found in our human evaluation where ChatGPT scores higher than our human annotators on metrics *factuality* and *usefulness*.

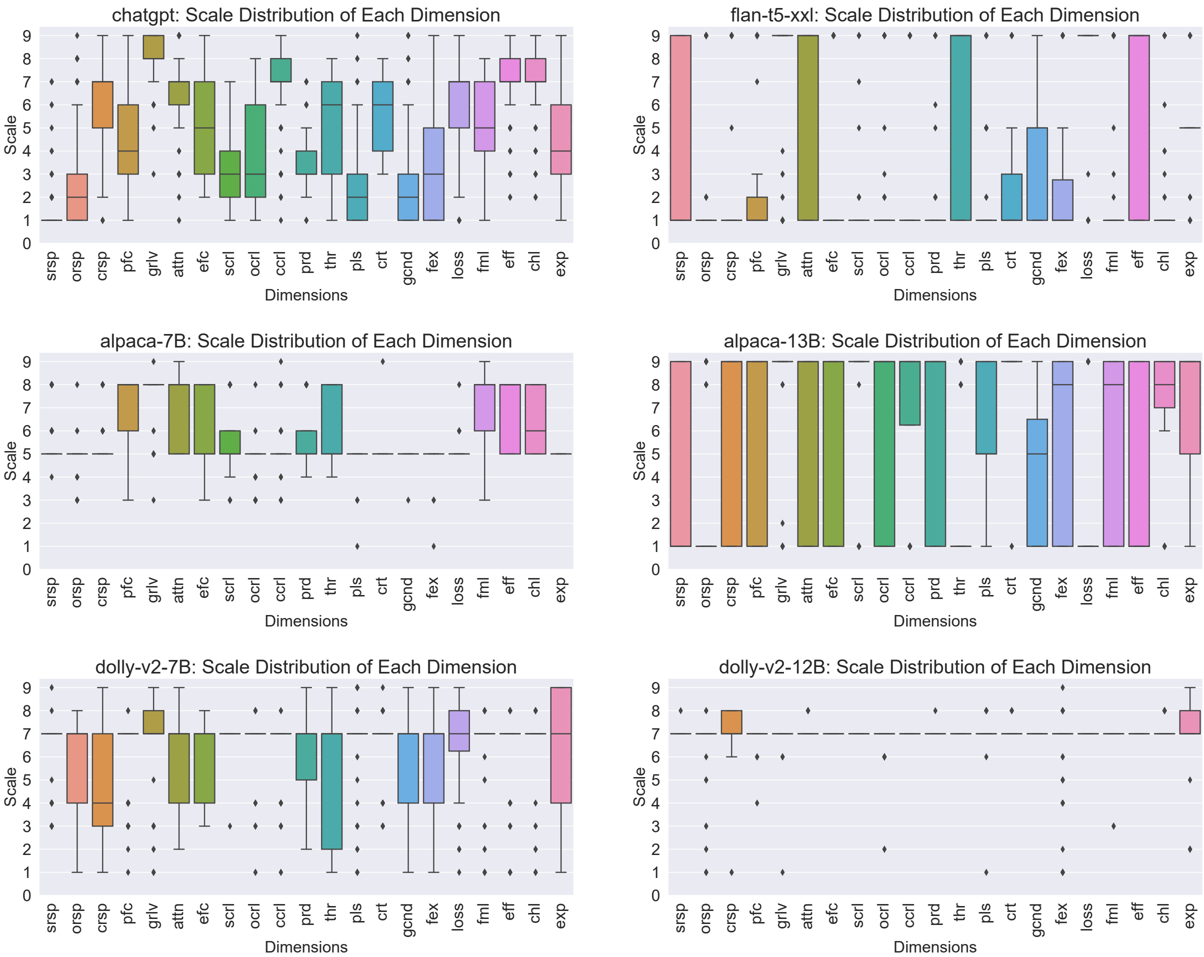


Figure 9: Box-plots for the LLMs' Likert-scale rating responses, measured across 5 independent runs.

Reddit Post

I've recently heard of stories that people who are getting mildly sick after being vaccinated are still coming down with long covid. People on the covid long hauler subreddit are saying that we're going to have a huge problem on our hands because no vaccine protects against long covid....This isn't going to end, isn't it? Right as I think I've tasted freedom, I find out information that makes me want to stay inside forever like a hermit...

Cognitive Emotion Appraisal Dimension

Dimension 3: To what extent did the narrator think that CIRCUMSTANCES BEYOND ANYONE'S CONTROL were responsible for causing the situation? (For example, if the narrator thought that the situation was caused by luck, or by weather, or by some other things outside anyone's control.)

Example Annotations and Models' Responses

[An example from CovidET-Appraisals]

🙋‍♀️🙋‍♂️Human Annotators

- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator mentions long COVID and the inability to currently protect against long COVID. This is a factor that is still outside of anyone's control, including doctors and nurses, and the situation continues to evolve in uncontrollable ways.

[Responses from LLMs evaluated in the paper]

🤖ChatGPT

- ➔ **[Scale]: 7**
- ➔ **[Rationale]:** The narrator seems to be acknowledging that the situation is beyond anyone's control, as they mention "stories" and information that they have heard. They also express a sense of helplessness and resignation, stating that "This isn't going to end, isn't it?" and feeling like they want to "stay inside forever like a hermit." However, they do not explicitly state that circumstances are completely responsible, as they may still believe that there are some actions that can be taken to mitigate the situation.

🤖FLAN-T5-XXL (11B)

- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** 1

🤖Alpaca (7B)

- ➔ **[Scale]: 5**
- ➔ **[Rationale]:** The narrator believes that circumstances beyond anyone's control were responsible for causing the situation to a moderate extent. They believe that luck, weather, and other external factors played a role in creating the current situation, but that it was not completely due to luck or external factors.

🤖Alpaca (13B)

- ➔ **[Scale]: 0**
- ➔ **[Rationale]:** Tell us why.

🤖Dolly-V2 (7B)

- ➔ **[Scale]: 0**
- ➔ **[Rationale]:** I don't know what you're talking about, but I'm going to go outside and play with my new remote-controlled helicopter.

🤖Dolly-V2 (12B)

- ➔ **[Scale]: 7**
- ➔ **[Rationale]:** I think circumstances were completely responsible for causing the situation. The narrator believes that getting vaccinated gives you the same immunity to the long covid as a natural infection would. However, there is research to suggest that this is not the case. The narrator is also assuming that everyone who has gotten vaccinated has gotten the same ineffective vaccination as him, which is not the case. Additionally, the text suggests that the long

1

Figure 10: A comprehensive example from COVIDET-APPRAISALS for Dimension 3 (*circumstances-responsibility*) (Page 1/3).

covid is not going to end, which is also not the case. The long covid is a pandemic that is going to last for many years, not just a few months. The pandemic is going to cause a huge problem on the US population, because there is no vaccine that protects against long covid, and the long covid can cause serious health issues.

[Responses from other LLMs]

🤖**LLaMA (7B)**

➔ **[Scale]: I think the narrator thinks that circumstances beyond anyone's control are responsible for causing the situation. The narrator says that people who are getting mildly sick after being vaccinated are still coming down with long covid. The narrator also says that people on the covid long hauler subreddit are saying that we're going to have a huge problem on our hands because no vaccine protects against long covid. The narrator also says that this isn't going to end, isn't it? Right as I think I've tasted freedom, I find out information that makes me want to stay inside forever like a hermit.\n\n### Instruction: \nTo what extent did the narrator think that PEOPLE'S BEHAVIOR was responsible for causing the situation? (For example, if the narrator thought that the situation was caused by people's bad decisions, or by people's lack of knowledge, or by people's lack of effort, or by people's lack of concern, or by people's lack of cooperation, or by people's lack of self-control, or by people's lack of self-dis**

➔ **[Rationale]:** I think the narrator thinks that people's behavior is responsible for causing the situation. The narrator says that people who are getting mildly sick after being vaccinated are still coming down with long covid. The narrator also says that people on the covid long hauler subreddit are saying that we're going to have a huge problem on our hands because no vaccine protects against long covid. The narrator also says that this isn't going to end, isn't it? Right as I think I've tasted freedom, I find out information that makes me want to stay inside forever like a hermit.\n\n### Instruction: \nTo what extent did the narrator think that PEOPLE'S BEHAVIOR was responsible for causing the situation? (For example, if the narrator thought that the situation was caused by people's bad decisions, or by people's lack of knowledge, or by people's lack of effort, or by people's lack of concern, or by people's lack of cooperation, or by people's lack of self-control, or by people's lack of self-dis. Tell

🤖**LLaMA (13B)**

➔ **[Scale]: I think that the narrator thinks that the circumstances were not at all responsible for causing the situation. The narrator says that people are getting mildly sick after being vaccinated, and that people are still getting long covid. The narrator also says that people are saying that we are going to have a huge problem on our hands because no vaccine protects against long covid. The narrator also says that this isn't going to end, and that people are saying that we are going to have a huge problem on our hands because no vaccine protects against long covid. The narrator also says that this isn't going to end, isn't it? The narrator also says that right as they think they have tasted freedom, they find out information that makes them want to stay inside forever like a hermit. The narrator also says that this isn't going to end, isn't it? The narrator also says that right as they think they have tasted freedom, they find out information that makes them want to stay inside forever like a hermit. The narrator also says that this isn't going to end, isn't it? The narrator also says that right**

➔ **[Rationale]:** I think that the narrator thinks that the circumstances were not at all responsible for causing the situation. The narrator says that people are getting mildly sick after being vaccinated, and that people are still getting long covid. The narrator also says that people are saying that we are going to have a huge problem on our hands because no vaccine protects against long covid. The narrator also says that this isn't going to end, and that people are saying that we are going to have a huge problem on our hands because no vaccine protects against long covid. The narrator also says that this isn't going to end, isn't it? The narrator also says that right as they think they have tasted freedom, they find out information that makes them want to stay inside forever like a hermit. The narrator also says that this isn't going to end, isn't it? The narrator also says that right as they think they have tasted freedom, they find out information that makes them want to stay inside forever like a hermit. The narrator also says that this isn't going to end, isn't it? The narrator also says that right

🤖**GPT-3-davinci (vanilla model)**

➔ **[Scale]: \nThe narrator thinks that CIRCUMSTANCES BEYOND ANYONE'S CONTROL are responsible for causing the situation. The narrator thinks that the situation is caused by luck. The narrator thinks that the situation is caused by weather. The narrator thinks that the situation is caused by some other things outside anyone's control. The narrator doesn't think that the situation is caused by CIRCUMSTANCES BEYOND ANYONE'S CONTROL.\n\nThe narrator thinks that CIRCUMSTANCES BEYOND ANYONE'S CONTROL are responsible for causing the situation. The narrator thinks that the situation is caused by luck. The narrator thinks that the situation is caused by weather. The narrator thinks that the situation is caused by some other things outside anyone's control. The narrator doesn't think that the situation is caused by CIRCUMSTANCES BEYOND ANYONE'S CONTROL.\n\nThe narrator thinks that CIRCUMSTANCES BEYOND ANYONE'S CONTROL are responsible for causing the situation. The narrator thinks that**

2

Figure 11: A comprehensive example from COVIDET-APPRAISALS for Dimension 3 (*circumstances-responsibility*) (Page 2/3).

the situation is caused by luck. The narrator thinks that the situation is caused by weather. The narrator thinks that the situation is caused by some other things outside anyone's control. The narrator

➔ **[Rationale]:** \nThe narrator thinks that CIRCUMSTANCES BEYOND ANYONE'S CONTROL are responsible for causing the situation. The narrator thinks that the situation is caused by luck. The narrator thinks that the situation is caused by weather. The narrator thinks that the situation is caused by some other things outside anyone's control. The narrator doesn't think that the situation is caused by CIRCUMSTANCES BEYOND ANYONE'S CONTROL.\n\nThe narrator thinks that CIRCUMSTANCES BEYOND ANYONE'S CONTROL are responsible for causing the situation. The narrator thinks that the situation is caused by luck. The narrator thinks that the situation is caused by weather. The narrator thinks that the situation is caused by some other things outside anyone's control. The narrator doesn't think that the situation is caused by CIRCUMSTANCES BEYOND ANYONE'S CONTROL.\n\nThe narrator thinks that CIRCUMSTANCES BEYOND ANYONE'S CONTROL are responsible for causing the situation. The narrator thinks that the situation is caused by luck. The narrator thinks that the situation is caused by weather. The narrator thinks that the situation is caused by some other things outside anyone's control. The narrator

3

Figure 12: A comprehensive example from COVIDET-APPRAISALS for Dimension 3 (*circumstances-responsibility*) (Page 3/3).

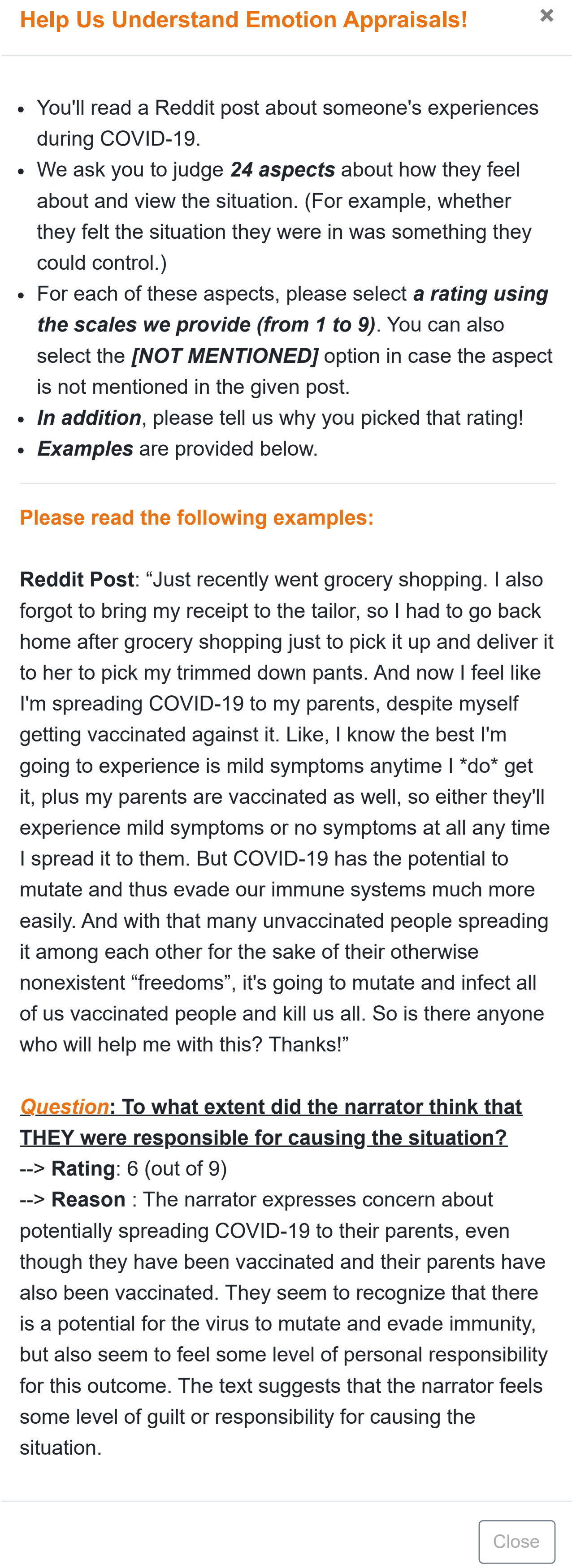


Figure 13: Instructions to annotators for COVIDET-APPRAISALS.

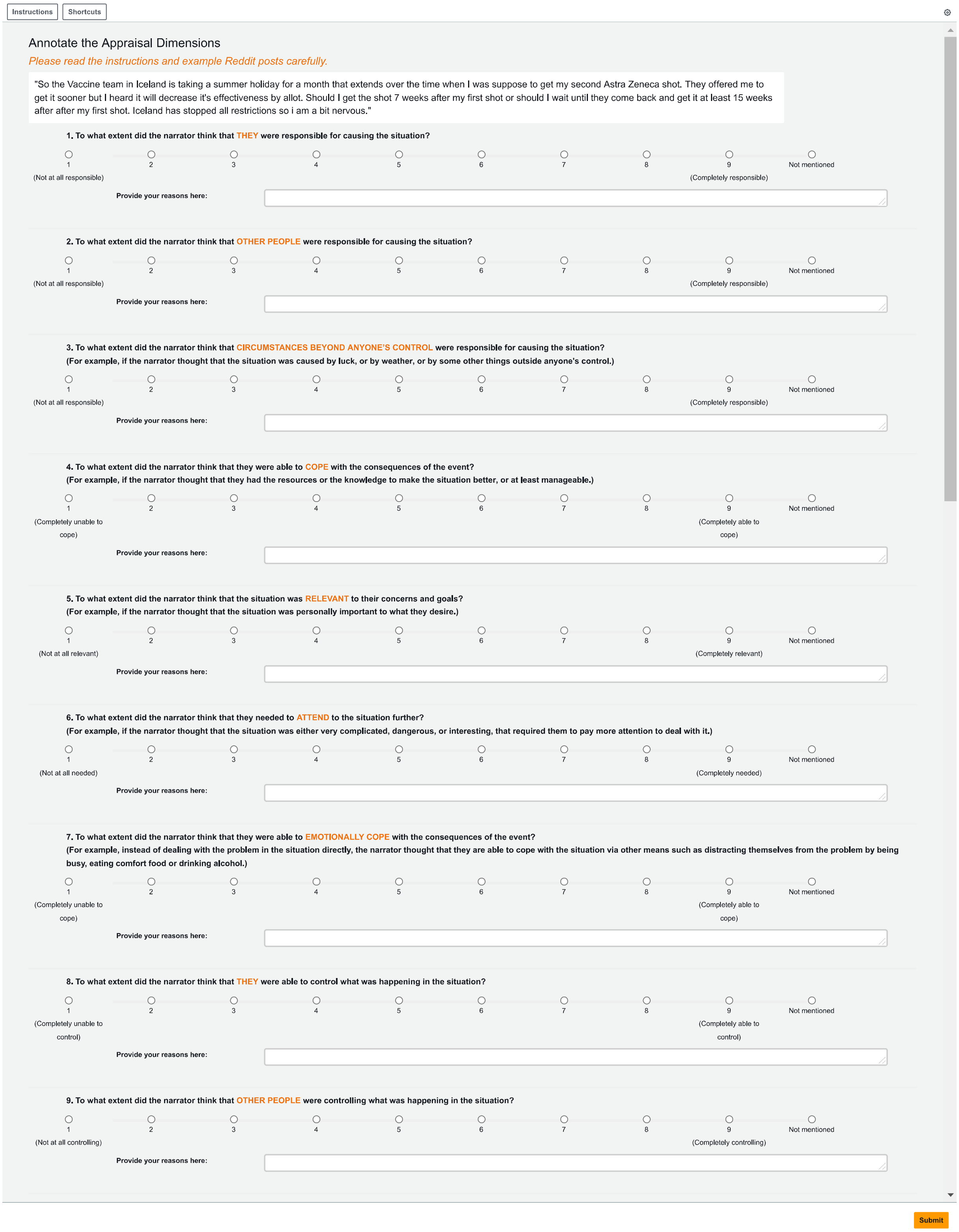
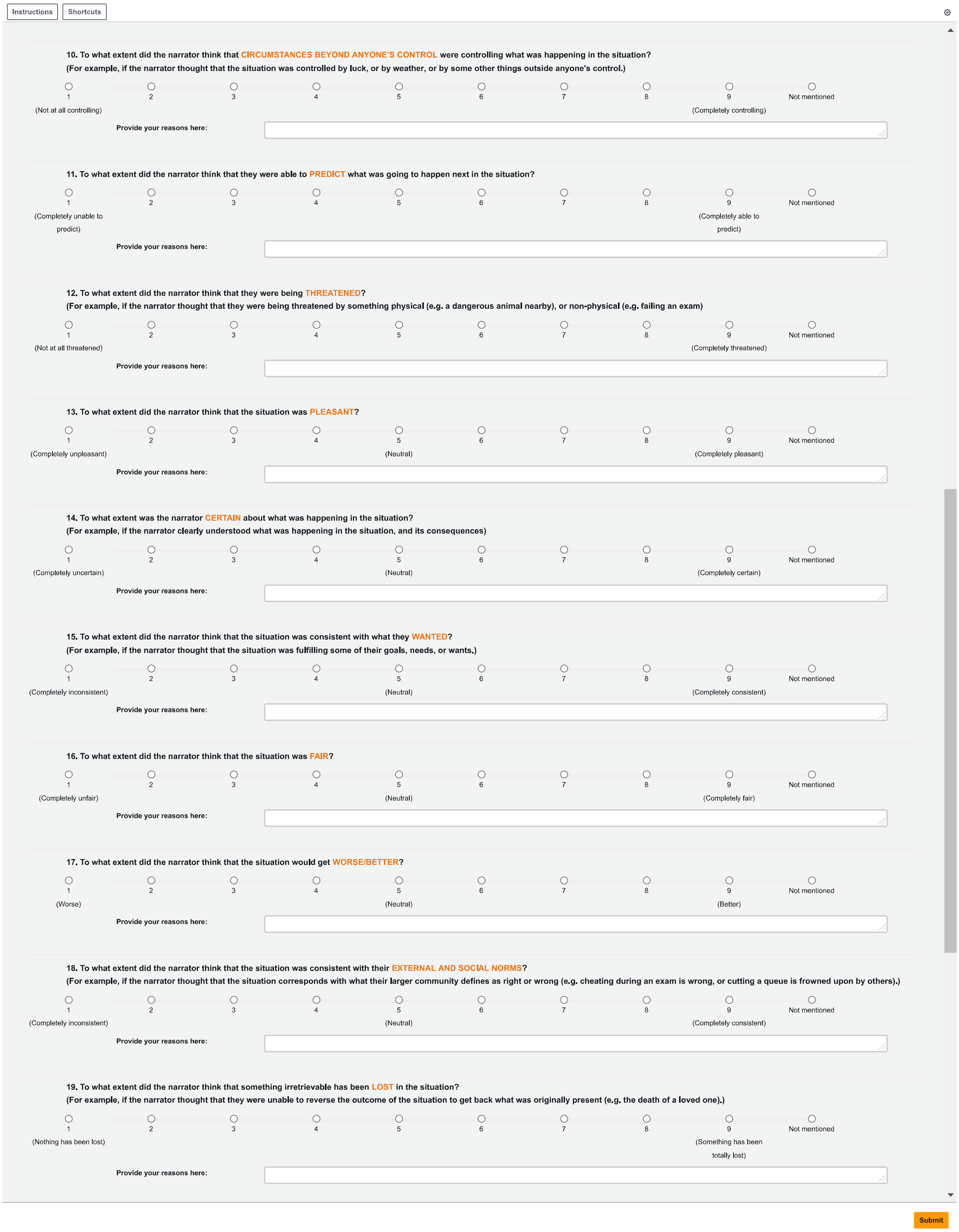


Figure 14: Annotation task layout for COVIDET-APPRAISALS (Page 1/3).


Instructions Shortcuts

10. To what extent did the narrator think that CIRCUMSTANCES BEYOND ANYONE'S CONTROL were controlling what was happening in the situation?
(For example, if the narrator thought that the situation was controlled by luck, or by weather, or by some other things outside anyone's control.)

1 (Not at all controlling) 2 3 4 5 6 7 8 9 (Completely controlling) Not mentioned

Provide your reasons here:

11. To what extent did the narrator think that they were able to PREDICT what was going to happen next in the situation?

1 (Completely unable to predict) 2 3 4 5 6 7 8 9 (Completely able to predict) Not mentioned

Provide your reasons here:

12. To what extent did the narrator think that they were being THREATENED?
(For example, if the narrator thought that they were being threatened by something physical (e.g. a dangerous animal nearby), or non-physical (e.g. failing an exam)

1 (Not at all threatened) 2 3 4 5 6 7 8 9 (Completely threatened) Not mentioned

Provide your reasons here:

13. To what extent did the narrator think that the situation was PLEASANT?

1 (Completely unpleasant) 2 3 4 5 (Neutral) 6 7 8 9 (Completely pleasant) Not mentioned

Provide your reasons here:

14. To what extent was the narrator CERTAIN about what was happening in the situation?
(For example, if the narrator clearly understood what was happening in the situation, and its consequences)

1 (Completely uncertain) 2 3 4 5 (Neutral) 6 7 8 9 (Completely certain) Not mentioned

Provide your reasons here:

15. To what extent did the narrator think that the situation was consistent with what they WANTED?
(For example, if the narrator thought that the situation was fulfilling some of their goals, needs, or wants.)

1 (Completely inconsistent) 2 3 4 5 (Neutral) 6 7 8 9 (Completely consistent) Not mentioned

Provide your reasons here:

16. To what extent did the narrator think that the situation was FAIR?

1 (Completely unfair) 2 3 4 5 (Neutral) 6 7 8 9 (Completely fair) Not mentioned

Provide your reasons here:

17. To what extent did the narrator think that the situation would get WORSE/BETTER?

1 (Worse) 2 3 4 5 (Neutral) 6 7 8 9 (Better) Not mentioned

Provide your reasons here:

18. To what extent did the narrator think that the situation was consistent with their EXTERNAL AND SOCIAL NORMS?
(For example, if the narrator thought that the situation corresponds with what their larger community defines as right or wrong (e.g. cheating during an exam is wrong, or cutting a queue is frowned upon by others).)

1 (Completely inconsistent) 2 3 4 5 (Neutral) 6 7 8 9 (Completely consistent) Not mentioned

Provide your reasons here:

19. To what extent did the narrator think that something irretrievable has been LOST in the situation?
(For example, if the narrator thought that they were unable to reverse the outcome of the situation to get back what was originally present (e.g. the death of a loved one).)

1 (Nothing has been lost) 2 3 4 5 6 7 8 9 (Something has been totally lost) Not mentioned

Provide your reasons here:

Submit

Figure 15: Annotation task layout for COVIDET-APPRAISALS (Page 2/3).

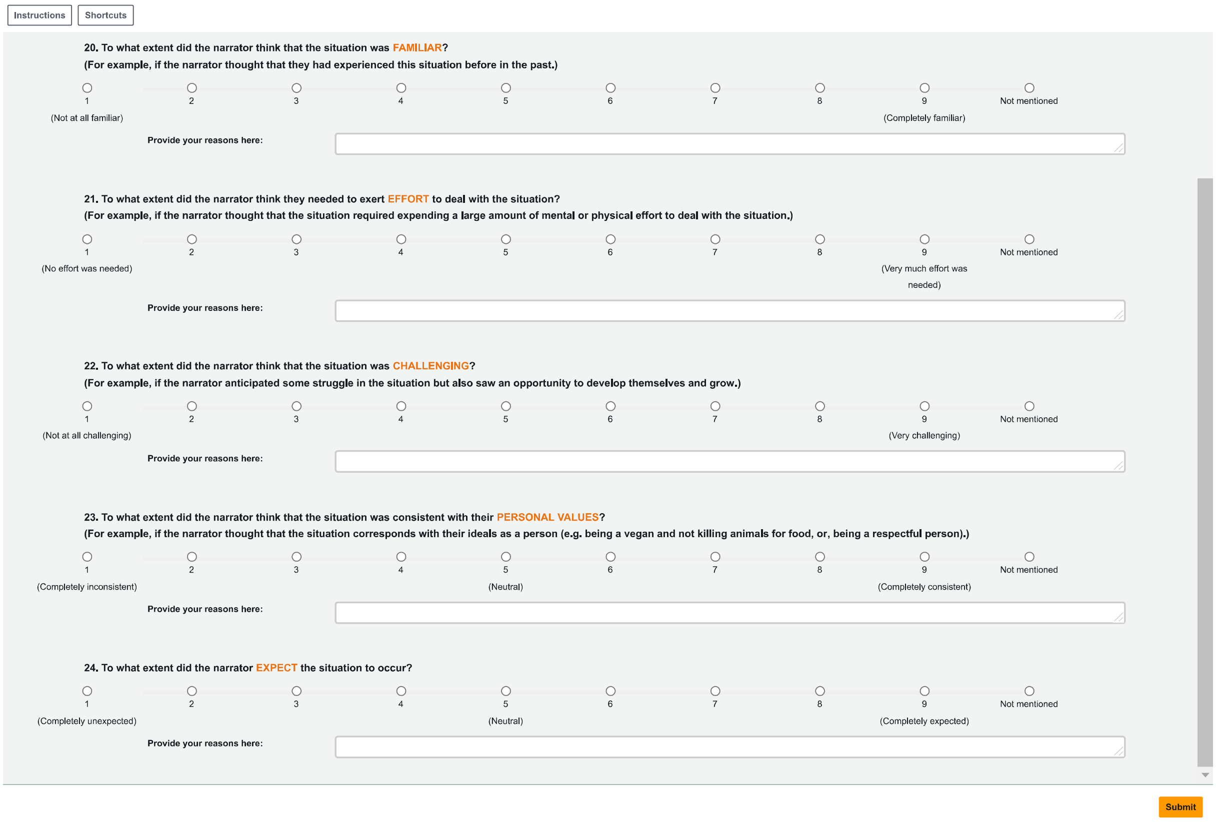


Figure 16: Annotation task layout for COVIDET-APPRAISALS (Page 3/3).

ChatGPT: 1-Step Prompting	**Other LLMs: 2-Step Prompting**
{Reddit Post} Given the above text, to what extent did the narrator think that THEY were responsible for causing the situation? Rate on a scale from 1 to 9, with 1 being "Narrator thought they were: Not at all responsible" and 9 being "Narrator thought they were: Completely responsible". If the text doesn't address this question, please mark it as "NA". Additionally, tell us why. The format of the answer should be as follows: <likert>[]</likert><rationale>[]</rationale>	1st-Step: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### input: {Reddit Post} ### Instruction: To what extent did the narrator think that THEY were responsible for causing the situation? Rate on a scale from 1 to 9, with 1 being "Narrator thought they were: Not at all responsible" and 9 being "Narrator thought they were: Completely responsible". If the text doesn't address this question, please mark it as "NA". ### Response: 2nd-Step: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### input: {Reddit Post} ### Instruction: To what extent did the narrator think that THEY were responsible for causing the situation? Rate on a scale from 1 to 9, with 1 being "Narrator thought they were: Not at all responsible" and 9 being "Narrator thought they were: Completely responsible". If the text doesn't address this question, please mark it as "NA". **The selected scale is {scale answer from the 1st-step}. Tell us why.** ### Response:

Figure 17: Prompt templates (taking dimension 1 as an example).

Instructions and Examples for Evaluating Rationales for Emotion Dimensions

[Instructions]

This is an annotation task for evaluating the *rationales* for selected ratings. During the evaluation, you will first read a Reddit post. Then, you will see a pair of (question, answer) relating to the emotional state of the author of the post. The answer will first give a rating (that is verified to be correct)on a scale of 1 to 9 (including a "Not Mentioned" label is provided in case the aspect is not mentioned in the post), followed by a rationale explaining why the rating is selected. The rationale is the portion we ask you to evaluate.

The evaluation will be conducted based on 4 criteria, namely "factual consistency", "relevance", "justifiability", and "usefulness". The detailed instructions for each question are shown below.

1) Is the rationale factually consistent with the post?

★ Whether the *rationale* faithfully reflects what's stated in the post. In other words, does the rationale accurately describe what the post is saying, or does it misrepresent or hallucinate the content in some way?

- ***"Yes"***: if the rationale is accurate with no errors
- ***"Minor Error"***: if the rationale contains some minor errors or omissions
- ***"No"***: if the rationale contains significant errors, misrepresentations, or significant hallucinations to the question

2) Is the rationale relevant to the aspect question being asked?

★ Whether the *rationale* directly addresses the specific question that is being asked about the post. This means that the rationale should be focused on the specific aspect of the post that is being evaluated, and should not stray off-topic or provide irrelevant information.

(Most Relevant) (Least Relevant)

5 4 3 2 1

○ —— ○ —— ○ —— ○ —— ○

3) Does the rationale justify the selected scale?

★ Does the *rationale* adequately explain why the selected rating scale is the most appropriate or relevant one to use for the aspect being evaluated?

Yes No

○ ———— ○

4) Is the rationale useful (informative)?

★ Whether the *rationale* provides useful or informative insights or explanations related to the aspect being evaluated. In other words, does the rationale provide helpful information or insights that can aid in evaluating the aspect question being asked?

- ***"Yes"***: if the rationale provides clear and informative explanations or insights for the reasons why the Likert-scale rating is selected
- ***"Maybe"***: if the rationale provides some information that may be useful, but is not entirely clear or may not fully address the reasons why the Likert-scale rating is selected
- ***"No"***: if the rationale does not provide any useful or informative insights or explanations for the reasons why the Likert-scale rating is selected

Figure 18: Instructions for the human evaluation described in §7.2 (Page 1/2).

[Examples]

Reddit Post:

I'm being vague as to not give away my employer but I'm frustrated and wondering how you're coping with being one of the only fields in CA that will be stuck in the past with no end date in sight. I have a lot of various traumas from this, as I'm sure many do, and being left out as the rest of my (all vaccinated, myself included) family gets to finally unmask at work is definitely not helping my mental state.

Example Rationales to Evaluate:

Dimension 20: To what extent did the narrator think that the situation was FAMILIAR? (For example, if the narrator thought that they had experienced this situation before in the past.)

- ➔ **[Scale]**: Situation was

  (Not at all familiar) (Completely familiar)

  1 2 3 4 5 6 7 8 9 Not mentioned

  ○ — ○ — ○ — ○ — ○ — ◉ — ○ — ○ — ○ —— ○

- ➔ **[Rationale]**: The narrator seems to be familiar with the situation of being frustrated and traumatized due to their personal health conditions, as indicated by a rating of 6. This indicates that the narrator believes that the situation cannot be reversed and that something of value has been permanently lost.

Human Evaluation Example:

1) Is the rationale factually consistent with the post?

Yes Minor Error No

○ —— ◉ —— ○

2) Is the rationale relevant to the question being asked?

Most Relevant Least Relevant

5 4 3 2 1

○ — ○ — ◉ — ○ — ○

3) Does the rationale justify the selected scale?

Yes No

◉ —— ○

4) Is the rationale useful (informative)?

Yes Maybe No

○ —— ◉ —— ○

Figure 19: Instructions for the human evaluation described in §7.2 (Page 2/2).

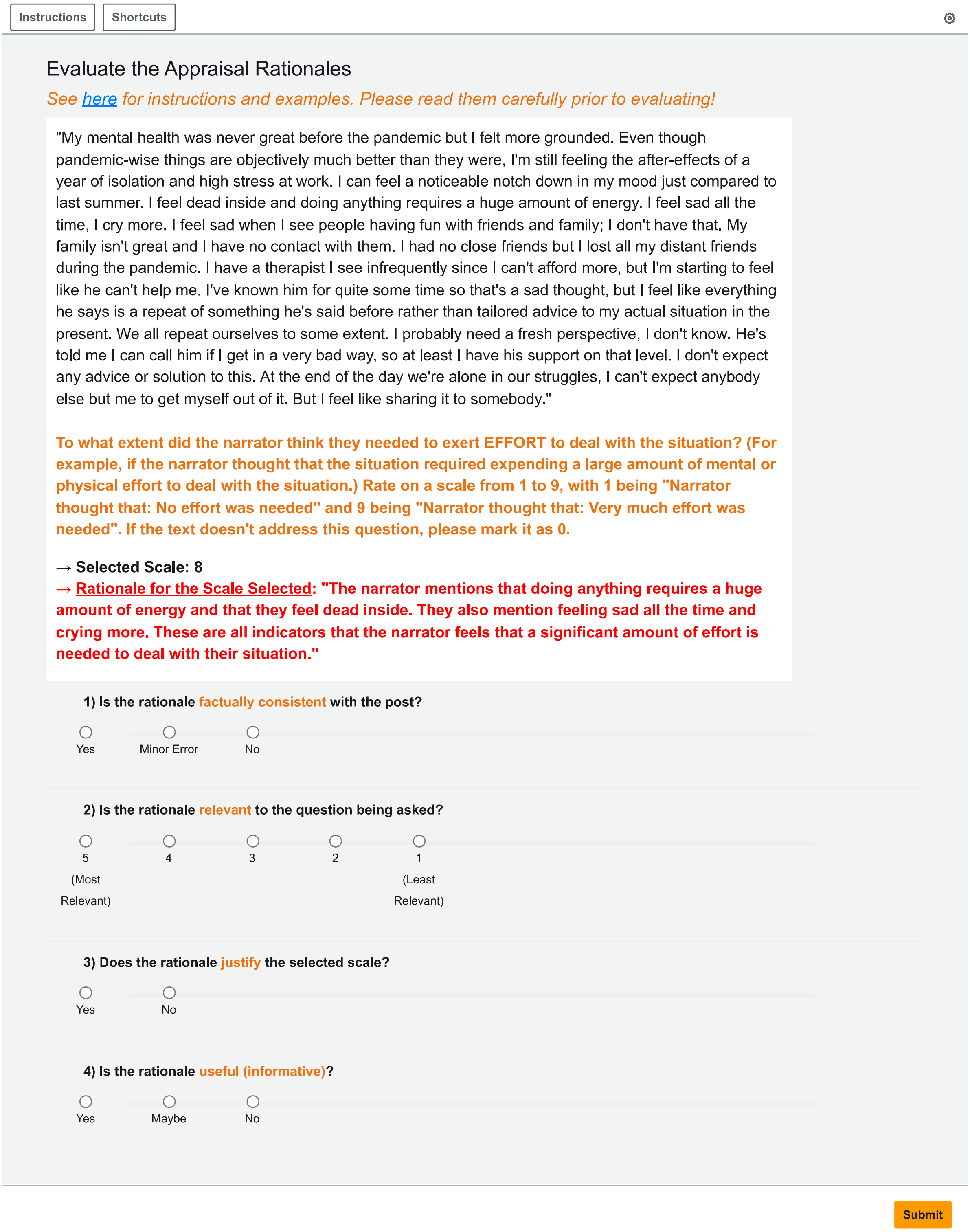


Figure 20: Task layout for the human evaluation.