Evaluating Subjective Appraisals of Emotions in Large Language Models

Anonymous EMNLP submission

Abstract

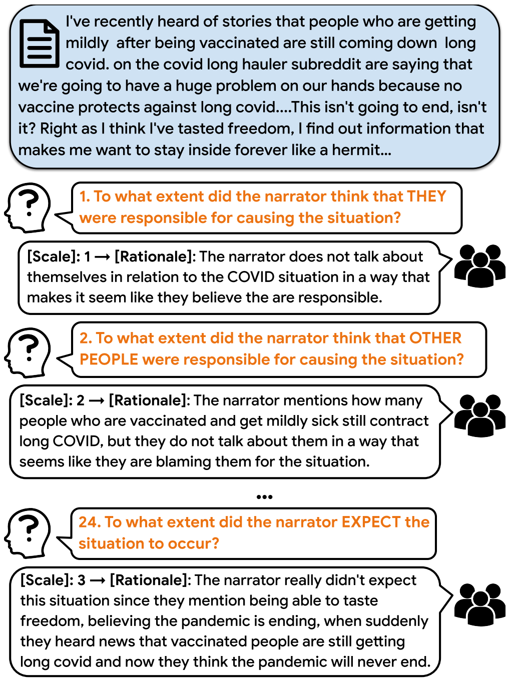
Emotions arise from the subjective cognitive appraisals that follow a situation. In this paper, we evaluate current Large Language Models' (LLM) ability to appraise subjective emotion dimensions (the sense-making that leads to emotions) from the text. We introduce a dataset entitled CovidET-Appraisals of ~200 Reddit posts annotated with 24 subjective cognitive appraisals that follow a situation as well as the rationales behind the appraisal dimension. CovidET-Appraisals presents an ideal testbed for commonsense emotional reasoning in LLMs.

Hongli: Need to update all figures pending new data

1 Introduction

[desmond: will modify cites] Emotions arise from the subjective cognitive appraisals that follow a situation (Ong et al., 2019). In this study, we propose a testbed entitled CovidET-Appraisals for evaluating Large Language Models' (LLM) ability in appraising emotions that take place under a certain situation. CovidET-Appraisals consists of Reddit posts annotated with Likert-scale ratings for 24 subjective emotion appraisal dimensions as well as the rationales for the rating selections. The rationales given in the prompt 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 showcase an example from CovidET-Appraisals in Figure 1. A more comprehensive example is demonstrated in Appendix §A.

Hongli: fill



I've recently heard of stories that people who are getting mildly after being vaccinated are still coming down long covid. 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...

1. To what extent did the narrator think that THEY were responsible for causing the situation?

[Scale]: 1 ⟶ [Rationale]: The narrator does not talk about themselves in relation to the COVID situation in a way that makes it seem like they believe the are responsible.

2. To what extent did the narrator think that OTHER PEOPLE were responsible for causing the situation?

[Scale]: 2 ⟶ [Rationale]: The narrator mentions how many people who are vaccinated and get mildly sick still contract long COVID, but they do not talk about them in a way that seems like they are blaming them for the situation.

...

24. To what extent did the narrator EXPECT the situation to occur?

[Scale]: 3 ⟶ [Rationale]: The narrator really didn't expect 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.

Figure 1: An example from CovidET-Appraisals (more details provided in Appendix §A).

Using CovidET-Appraisals, we explore the following research questions. From a philosophical standpoint, we ask **1)** Whether subjective appraisals change emotions (e.g., would the narrator have different emotions when they feel controllable or out of control for the situation); and **2)** Can LLMs understand such appraisals? From an evaluation point of view, we explore **3)** The feasibility of automatic evaluation metrics in assessing the quality of rationales for subjective emotion appraisals.

Hongli: We release a benchmark with metrics. (Cite Gerhmann talk stuff) Gerhmann talk: (Howcroft et al., 2020; Yuan et al., 2022; Gehrmann et al., 2022a)

We evaluate the zero-shot performance of LLMs in an attempt to evaluate their innate emotion appraisal ability.

"A new paradigm for metrics What if, instead of relying on existing metrics, a benchmark can be released with its own metrics? We are saving a ton by not needing large training corpora. So let's collect human annotations as metric training data. Annotate validation outputs to train metrics, and test outputs to evaluate systems AND the new metrics" (Gehrmann et al., 2022b).

2 Background and Related Work

Evaluating LLMs. Zero-shot elicitation is more suitable for understanding the innate capability possessed by the LLMs.

ID	Appraisal Dimensions
1	*Self-responsibility*
2	*Other-responsibility*
3	*Circumstances-responsibility*
4	*Problem-focused coping*
5	*Goal Relevance*
6	*Attentional activity*
7	*Emotion-focused coping*
8	*Self-Controllable*
9	*Other-Controllable*
10	*Circumstances-Controllable*
11	*Predictability*
12	*Threat*
13	*Pleasantness*
14	*Certainty*
15	*Goal Conduciveness*
16	*Fairness*
17	*Future expectancy*
18	*Consistency with social norms*
19	*Loss*
20	*Familiarity*
21	*Effort*
22	*Challenge*
23	*Consistency with internal values*
24	*Expectedness*

Table 1: The 24 appraisal dimensions we used. See Figure 1 for an example of how the items for 1: *self-responsibility*, 2: *other-responsibility*, and 24: *expectedness* were framed.

Positive Re-framing. (Ziems et al., 2022)

Scene and Story-Telling. The different dimensions in this work can be interpreted instead as different ways of telling the same story.

Hongli: Cite Katrin's github repo.

3 Dataset Construction

We present CovidET-Appraisals, a dataset of English Reddit posts annotated with perceived subjective appraisals of the situation that the narrator of the post is undergoing.

3.1 Data Source

We use Reddit posts curated in CovidET[1] (Zhan et al., 2022) as the source of our data. CovidET includes $1,883$ Reddit posts sourced from `r/COVID19_support`[2] during COVID-19. Each post is manually annotated with one or many of the 7 emotions (namely *anger*, *anticipation*, *joy*, *trust*, *fear*, *sadness*, and *disgust*). The posts in CovidET are lengthy and emotionally rich, with an average of 156.4 tokens and 2.46 emotions per post. For the annotation of CovidET-Appraisals, we sample posts from CovidET.

Hongli: fill

3.2 Annotation Task

We collect annotations on the Amazon Mechanical Turk (MTurk). We follow the annotation protocols used in the Stanford Emotional Narratives Dataset [desmond: TO FIX THIS PART, IT'S A NEW PROTOCOL] (Yeo and Ong, 2023). We provide the instructions given to the annotators as well as the layout of the annotation task in Appendix §B.

Annotators. We recruit 2 linguistic students at a university to work on our annotation task[3]. Both annotators were involved in a pre-annotation *qualification as well as training* process before commencing annotating CovidET-Appraisals. The qualification and training process is described in Appendix §C. During the annotation, we consistently evaluate the inter-annotator agreement and provide feedback on the annotators' work.

Annotation 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 scales of 1 to 9. A "Not mentioned" option is also provided in case the appraisal dimension being asked is not present in the given post. In addition to selecting the Likert-scale ratings for each appraisal dimension in the situation, we also ask the annotators to provide the rationales for their rating selections. We provide the user-friendly labels for each appraisal dimension in Table 1. A comprehensive list of appraisal dimension questions is also provided in Figure 17, Figure 18, and Figure 19 in Appendix §B.

Owing to the subjective nature of the annotation task as well as the immense efforts and expertise inherently required for the annotation of the rationales[4], we assign 1 expert annotator per post for the annotation of the CovidET-Appraisals. To ensure the quality of the annotations, we randomly sample 40 posts and have them doubly annotated

---

[1] `https://github.com/honglizhan/CovidET`

[2] `https://www.reddit.com/r/COVID19_support/`

[3] Our group has a continuing working relationship with these linguists, and they possess rich prior experience in performing intricate tasks that are related to our current task.

[4] Our statistics show that even experienced expert annotators spend on average around 30 minutes to complete the annotation of one post, which includes the selection of the ratings for each of the 24 appraisal dimensions as well as the provision of the rationales.

by both expert linguists (see Section §4 for more details).

3.3 Benchmark Testbed

Hongli: fill

We collect posts annotated with Likert-scale ratings for the 24 emotion appraisal dimensions as well as the rationales for the rating selections. We convert the value of the "Not mentioned" label to 0, as we find during the annotation collection that the "Not mentioned" option tends to co-occur with lower scale ratings such as 1 and 2.

Label Aggregation. For the subset of 40 posts that are doubly annotated for quality measurement purposes, we aggregate the Likert-scale ratings by taking the mean of each post's ratings for each appraisal dimension. In terms of the rationales, we consider both rationales as ground truth references and use multi-reference to compute our metrics during experimental evaluations.

4 Inter-Annotator Agreement

In this section, we report the agreement among our expert linguist annotators for COVIDET-APPRAISALS. As previously mentioned, to secure the quality of COVIDET-APPRAISALS, we randomly sample 40 posts and have them doubly annotated by both expert annotators. In Table 2, we provide the inter-annotator agreement statistics between the annotators for both Likert-scale ratings as well as the rationales. The full report which includes a comprehensive overview of other metrics of the inter-annotator agreement statistics for the rationales in shown Appendix §D.

4.1 Likert-Scale Ratings

We first measure the inter-annotator agreement of the Likert-scale ratings using the Pearson correlation coefficient (Pearson's r). As mentioned in Section §3, the "Not mentioned" label is converted into an integer value of 0 as it tends to co-occur with lower-scale ratings. As shown in Table 2, both expert linguist annotators have significantly substantial agreement among themselves ($p < 0.1$), with the value of Pearson's r at 0.627.

Hongli: Update and fix, explain how the correlation is calculated.

In addition, we also measure the spread of the Likert-scale ratings within each dimension for each post using measures including the mean absolute delta as well as the standard deviation. Results show that the expert linguists yield small differences between their Likert-scale ratings, with a mean absolute delta of 1.905 and a standard deviation of 0.953. Furthermore, to measure the agreement among annotators for selecting the "Not mentioned" label, we average across dimensions the Fleiss' Kappa values (Fleiss, 1971; Randolph, 2005). [desmond: we don't need to cite Fleiss or the Randolph paper...] Both expert linguists have a considerably high agreement, with a Kappa of 0.769. The results of these metrics demonstrate strong accordance among our annotators in terms of Likert-scale rating selections.

Sub-Group Agreement. In the study conducted by Ong et al. (2019), the 24 emotion appraisal dimensions were divided into 4 groups. To better understand the inter-annotator agreement pertaining to each emotion appraisal dimension in COVIDET-APPRAISALS, we divide the 24 dimensions into 4 groups following (Ong et al., 2019). Specifically, each group has 6 appraisal dimensions in ascending order (e.g., group 1 has dimensions $1 \sim 6$, group 2 has dimensions $7 \sim 12$, group 3 has dimensions $13 \sim 18$, and group 4 has dimensions $19 \sim 24$). We provide the inter-annotator agreement statistics per group in Figure 2. As the plot shows, the expert annotators have a slightly lower agreement in dimensions 7 to 12, as well as dimensions 19 through 24.

4.2 Natural Language Rationales

To calculate the agreement between the expert linguist annotators on their rationale annotations, we apply automatic evaluation measures such as BLEU, ROUGE, and BERTScore. Specifically, we calculate the BLEU score with smoothing methods 1, and report the re-scaled BERTScore values in Table 2. We establish 2 random baselines for comparison, namely 1) we randomly sample 3 distinct rationales of the same dimension from different posts as candidates, and compute the metrics against ground truth references (i.e., same dimension, *different* posts); and 2) we randomly sample 3 different rationales from different dimensions within the same post as candidates, and compute the metrics against the ground truth references (i.e., same post, *different* dimensions).

As shown in Table 2, the expert linguists have a lower agreement in terms of word overlaps (BLEU-4 of 0.046, ROUGE-L of 0.262) compared to the random baselines. However, the agreement in terms of textual semantic similarity (BERTSC of 0.379) is (slightly) higher than the random base-

	SCALE				RATIONALE		
	ABS. DELTA	STD. DEV.	PEARSON CORR.	“NA” KAPPA	BLEU-4	ROUGE-L	BERTSC
EXPERTS	1.905	0.953	0.370 ($p = 0.122$)	0.769	0.046	0.262	0.379
– RATIONALE BASELINE (same dimension, *different posts*)					0.076	0.279	0.364
– RATIONALE BASELINE (same post, *different dimensions*)					0.059	0.247	0.337

Table 2: Inter-annotator statistics of COVIDET-APPRAISALS. More details are provided in Appendix D.

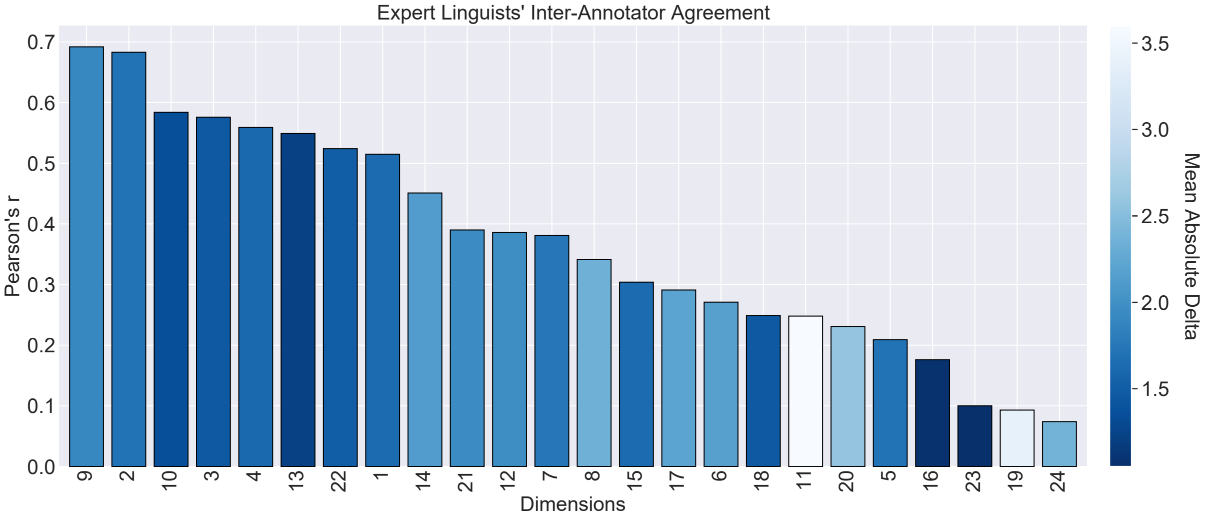


Figure 2: Pearson correlation and mean absolute delta of Likert-scale ratings within each group of dimensions, measured for the 40 posts doubly annotated by both expert linguist annotators.

lines. This demonstrates the incompetence of these automatic measures to evaluate such a subjective task as providing rationales for emotion appraisal dimensions. In particular, the word-overlap metrics do not capture the intrinsic subjective nature of the rationales. Taking the post in Figure 1 for example. With respect to the first appraisal dimension, 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 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. As a result, the automatic evaluation metrics thus fail to reflect the correctness of a rationale for a subjective emotion appraisal dimension.

5 Dataset Analysis

In this section, we analyze the subjective emotion appraisal dimensions as well as their rationales annotated in COVIDET-APPRAISALS. We provide the Z-score normalized analysis of the ratings in Appendix §E[5].

Dimension Distribution. The distribution of the Likert-scale ratings is shown in Figure 3. Additional analyses are provided in Appendix §E.

Hongli: Figures need update

Emotions and Dimensions Distribution. We analyze the connections between COVIDET-APPRAISALS’s Likert-scale annotations and COVIDET’s emotion annotations. Specifically, we measure the mean Likert-scale rating for each dimension within each post with respect to the perceived emotion. The emotion dimension distribution is shown in Figure 4.

Hongli: Figures need update

---

[5] In practice, we don’t apply normalization to analyze the dataset since in some dimensions the ratings tend to be consistent. Normalizing for each annotator within all dimensions would thus imply breaking the consistency of such dimensions.

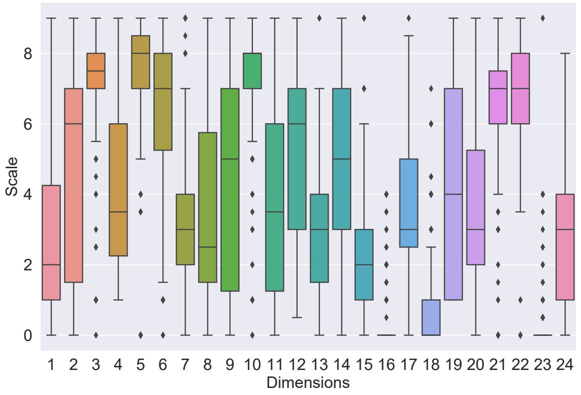


Figure 3: Box plot of the Likert-scale ratings for each dimension.

Predicting Emotions from Scales. We use a multi-class logistic regression model to predict the 7 classes of emotions using the scales of the annotated 24 appraisal dimensions as features. Specifically, we use the Likert-scale ratings annotated in the validation set as training data to predict the ratings in the test set. To handle the class imbalance issue, we use class weights during training, which gives more importance to the minority class by assigning a higher weight to its samples, while assigning a lower weight to the majority class samples.

The weights of the dimensions in the trained logistic regression model are in Figure 5.

Hongli: correlation with the logistic regression results: how they correlate with desmonds' psy-arxiv hypothesis with emotion / dimension paper.

Hongli: Note: no weight is significant.

Topic Variations in Rationales. To better understand the rationales for each dimension, we use the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to extract the topics in the rationales. Stopwords such as common English function words and words that occur frequently in our prompts (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 Figure 6.

6 Evaluating LLMs

We evaluate the zero-shot performance of LLMs in an attempt to evaluate their innate emotion appraisal ability. In practice, we find that non-instruction-tuned models have a hard time inferring from our prompts.

Hongli: Maybe show examples.

Models Under Inference. We evaluate the following instruction-tuned LLMs: ChatGPT, Alpaca (7B)[6], Dolly (7B)[7], and FLAN-T5 (XXL). ChatGPT uses PPO with learned reward models[8].

6.1 Experimental Setup

To avoid randomness and increase the stability of the models, we set the temperature for all models at 0.1[9].

Prompts. Due to recency bias discovered in LLMs , we place the appraisal question at the end of the prompt. The templates for prompting the LLMs are shown in Figure 7. For ChatGPT, we use one-step prompting. On the other hand, for all other models, we apply two-step prompting[10]: 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. Our method of two-step prompting is similar to "self-asking", where we ask the model follow-up questions needed here (Press et al., 2022).

Hongli: cite

To evaluate the performance, we clean the responses elicited from the LLMs[11].

6.2 Results and Discussions

The results of the LLM responses are shown in Table 3. A report of a more holistic overview of results from the automatic metrics is provided in Appendix §G. We report the average performance across five independent runs. We further analyze the LLMs' responses in Appendix §H.

Hongli: Pending

Examples of the models' responses are shown in Appendix §A and Appendix §F.

Automatic Evaluations. Models' scales and rationales: several things needed to improve them, such as self-consistency.

Hongli: Greg's fact checking repo.

7 Human Evaluation for Rationales

As previously discussed, the automatic metrics such as BLEU, ROUGE, and BERTScore fail to

---

[6] https://github.com/tatsu-lab/stanford_alpaca. Note that the max context length of Alpaca is 512.

[7] https://github.com/databrickslabs/dolly

[8] https://beta.openai.com/docs/model-index-for-researchers

[9] We experimented with higher temperatures, and they yielded worse and more unstable performance.

[10] We experimented with the other models using the 1-step prompt used in ChatGPT, and could not elicit well-formatted responses.

[11] For example, one of Alpaca's 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"*.

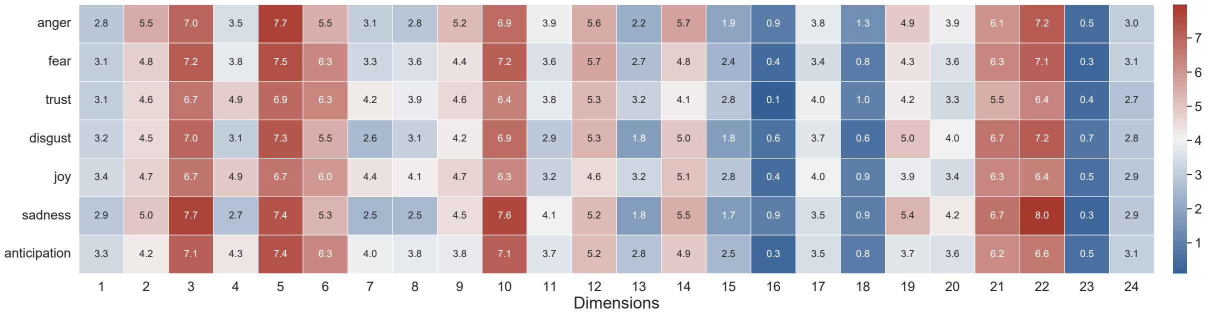


Figure 4: Mean Likert-scale ratings for each dimension in each emotion.

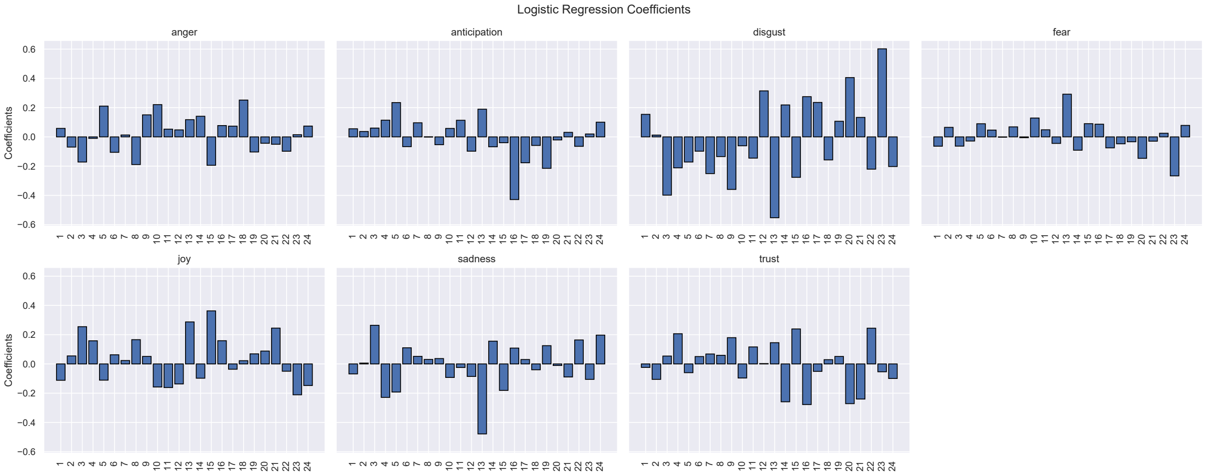


Figure 5: Weights of the trained logistic regression model.

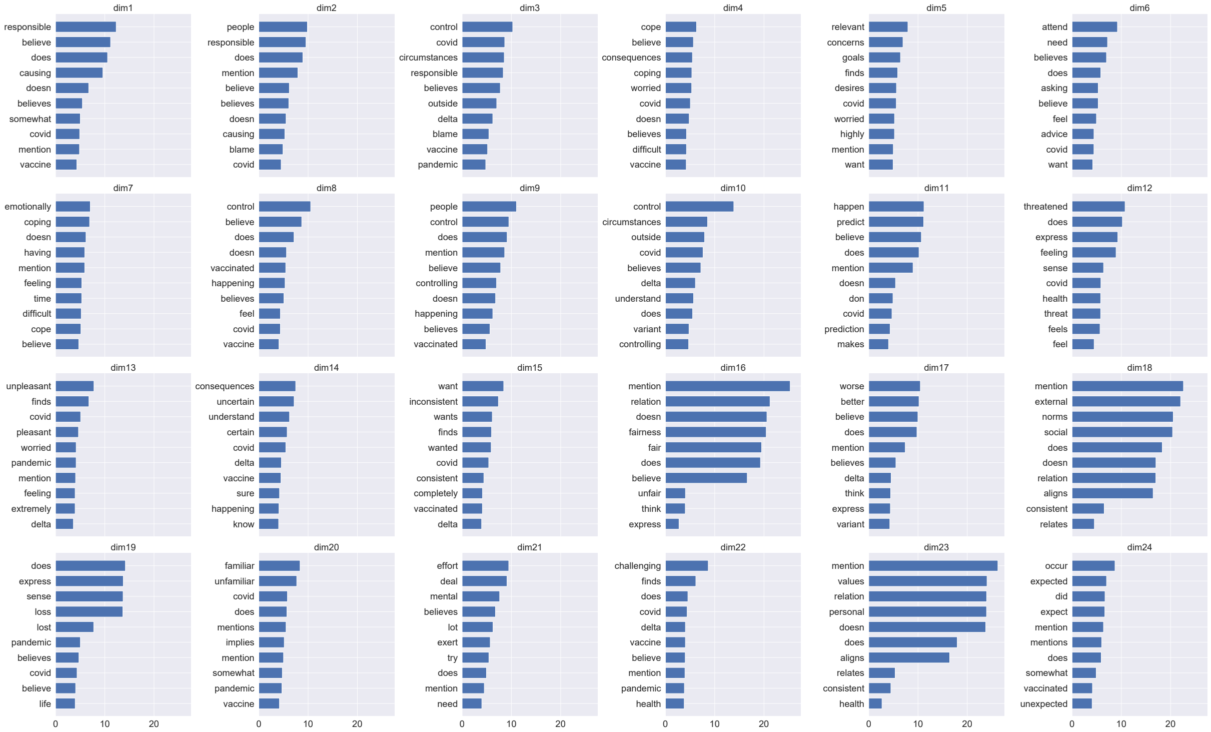


Figure 6: LDA results on the annotated rationales for each appraisal dimension.

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

faithfully reflect the quality of the rationales for the current task of providing rationales for subjective emotion appraisals in the situation. Therefore, in this section, we turn to human evaluation.

7.1 Evaluation Framework

We assess the rationales based on the following criteria:

1) Factuality. For the rationale, the model may not generate something that is factual: sometimes it generates rationales to justify its answer (Ye and Durrett, 2022). Therefore, we include the aspect of *hallucination and factuality* as one of our evaluation criteria.

2) Relevance. We evaluate whether the rationale provided is relevant to the aspect question being asked.

3) Consistency. We ask human evaluators whether the rationale justifies the selected scale. (the rationale it generates does not really agree with the output).

4) Usefulness. We ask whether the rationale provides useful information pertaining to the appraisal dimension.

7.2 Evaluation Task

We collect human evaluations from the Amazon Mechanical Turk (MTurk). We recruit MTurk crowd workers to work on our human evaluation task. The crowd workers are involved in a *qualification and training* process prior to commencing the annotations.

8 Conclusion

We present COVIDET-APPRAISALS.

This work presents the first step for understanding subjective emotion appraisals from text in today's LLMs.

Limitations

Future work can collect more data.

References

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

	Scale			Rationale		
	Abs. Delta	Pearson Corr.	"NA" Kappa	BLEU-4	ROUGE-L	BERTSc
ChatGPT	2.061	0.500 ($p = 0.124$)	0.723			
Alpaca-7B	2.808	0.287 ($p = 0.242$)	0.724			
Alpaca-13B	4.079	-0.014 ($p = 0.242$)	0.724			
Dolly-7B	3.857	-0.102 ($p = 0.494$)	0.019			
Dolly-12B	3.857	-0.102 ($p = 0.494$)	0.019			
Flan-T5-XXL	3.342	0.108 ($p = 0.352$)	0.555			

Table 3: Experiment results from LLMs. We report the average performance across five independent runs.

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

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022a. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

Sebastian Gehrmann, Sebastian Ruder, Vitaly Nikolaev, Jan A. Botha, Michael Chavinda, Ankur Parikh, and Clara Rivera. 2022b. Tata: A multilingual table-to-text dataset for african languages.

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.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning.

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

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2022. Synthbio: A case study in human-ai collaborative curation of text datasets.

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.

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

We showcase a comprehensive example from CovidET-Appraisals in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13, Figure 14, and Figure 15.

B Dataset Annotation Framework

We provide the instructions given to the annotators in Figure 16. In addition, we also provide the layout for the annotation task in Figure 17, Figure 18, and Figure 19.

C Annotator Qualification and Training

We train and qualify our annotators before allowing them to annotate on CovidET-Appraisals. During the *pre-qualification* process, annotators were given a set of posts to annotate and were subsequently compared with gold annotations from our group.

Hongli: Revise

D Full Inter-Annotator Statistics

The full statistics of the rationale agreement between the expert linguistics annotators for CovidET-Appraisals is shown in Table 4.

E Additional Dataset Analyses

E.1 Likert-Scale Rating Distribution Plots

We provide additional analyses of the distribution of Likert-scale ratings annotated in CovidET-Appraisals in Figure 20 and Figure 21.

E.2 Normalized Dataset Analysis

Here we show the analysis for the Z-score normalized version of the Likert-scale ratings in CovidET-Appraisals.

Rating Normalization. We normalize the Likert-scale rating annotations within each annotator. Specifically, we transform each annotator's Likert-scale annotations into a Z-scored value by first computing the mean and standard deviation for each annotator's ratings for all dimensions, then subtracting the mean from each rating and dividing by the standard deviation. After normalization, each annotator's ratings have a mean of 0 and a standard deviation of 1.

Likert-Scale Ratings. In Figure 22 and Figure 23, we provide the distribution of the annotators' Likert-scale ratings *pre-normalization*.

Hongli: Update

F Additional Model Responses

We also tried to elicit the responses from other models such as LLaMA[12] (Touvron et al., 2023) using the same template given in Figure 7. The additional responses from these models are shown in Figure 24 and Figure 25.

G Full LLM Rationale Measures

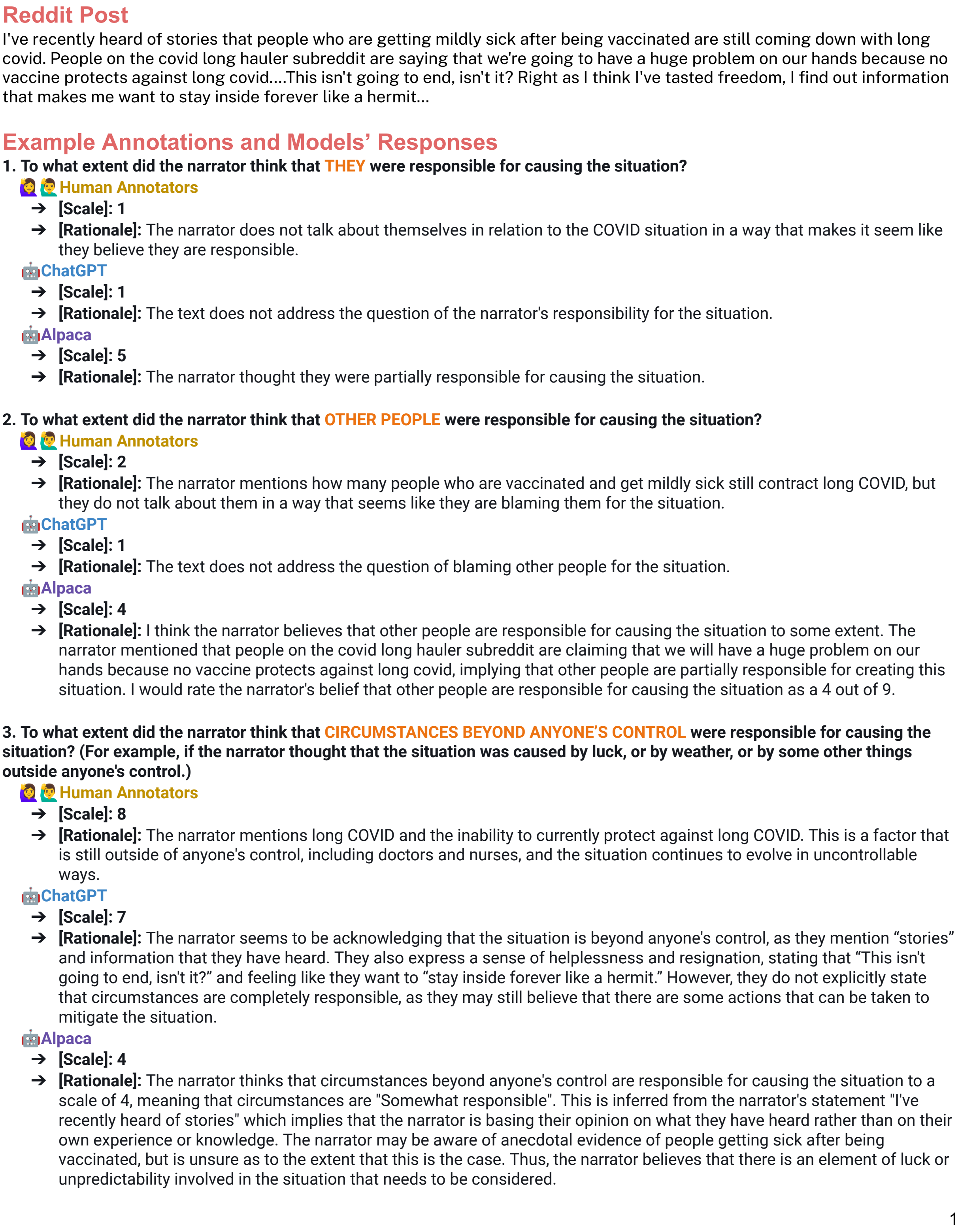
We provide the full statistics of the rationale agreement measured for the LLMs' responses against the gold annotations in Table 5.

H Model Responses Analyses

The LLMs' performance in terms of Pearson correlation and mean absolute delta against the gold annotations are shown in Figure 26 and Figure 27. Additionally, the box plots for each LLM's Likert-scale ratings are shown in Figure 28 and Figure 29.

---

[12] `https://github.com/facebookresearch/llama`



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...

Example Annotations and Models' Responses

1. To what extent did the narrator think that THEY were responsible for causing the situation?

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** The narrator does not talk about themselves in relation to the COVID situation in a way that makes it seem like they believe they are responsible.

🤖**ChatGPT**
- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** The text does not address the question of the narrator's responsibility for the situation.

🤖**Alpaca**
- ➔ **[Scale]: 5**
- ➔ **[Rationale]:** The narrator thought they were partially responsible for causing the situation.

2. To what extent did the narrator think that OTHER PEOPLE were responsible for causing the situation?

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 2**
- ➔ **[Rationale]:** The narrator mentions how many people who are vaccinated and get mildly sick still contract long COVID, but they do not talk about them in a way that seems like they are blaming them for the situation.

🤖**ChatGPT**
- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** The text does not address the question of blaming other people for the situation.

🤖**Alpaca**
- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** I think the narrator believes that other people are responsible for causing the situation to some extent. The narrator mentioned that people on the covid long hauler subreddit are claiming that we will have a huge problem on our hands because no vaccine protects against long covid, implying that other people are partially responsible for creating this situation. I would rate the narrator's belief that other people are responsible for causing the situation as a 4 out of 9.

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.)

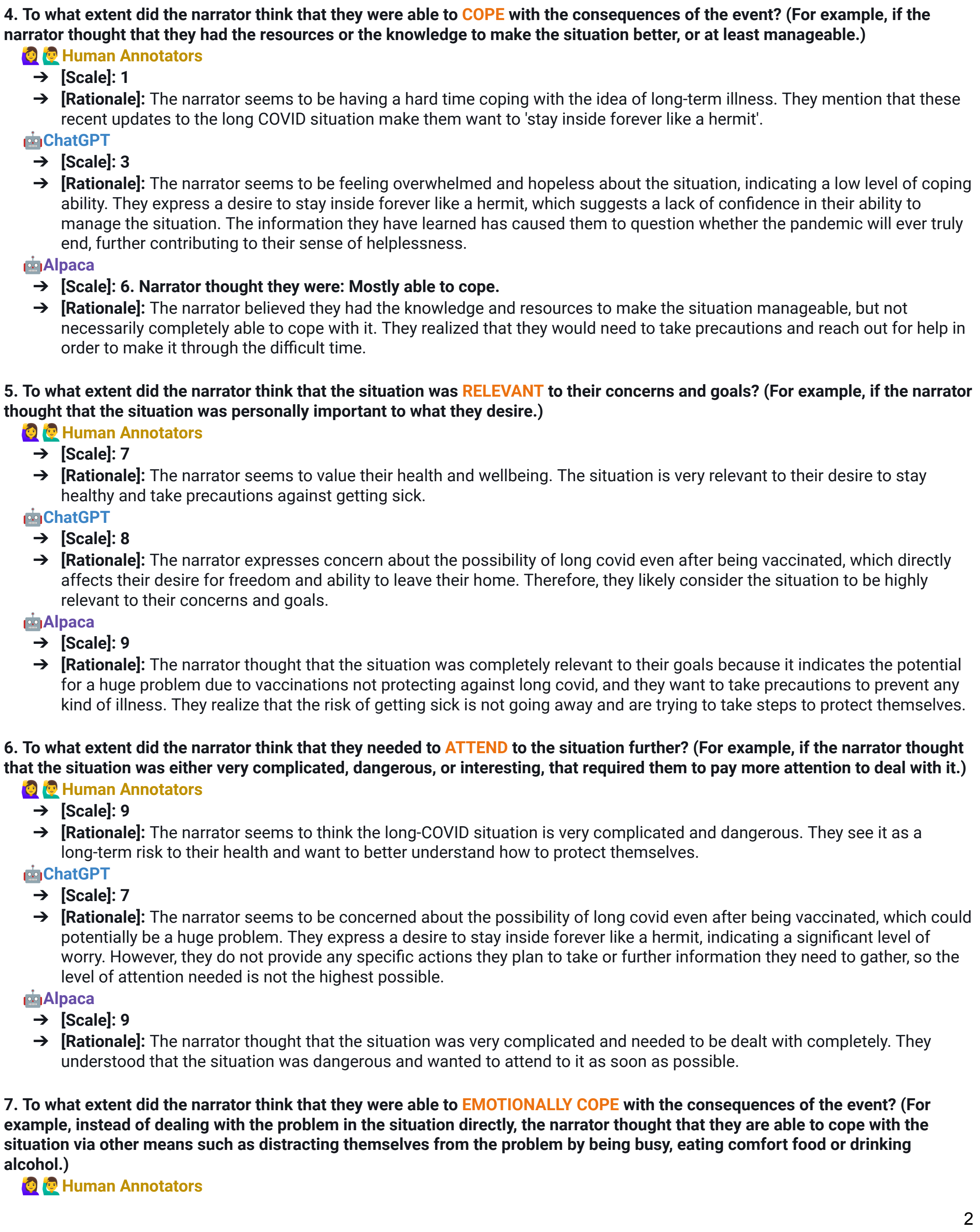
🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator mentions long COVID and the inability to currently protect against long COVID. This is a factor that is still outside of anyone's control, including doctors and nurses, and the situation continues to evolve in uncontrollable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

🤖**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.

🤖**Alpaca**
- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** The narrator thinks that circumstances beyond anyone's control are responsible for causing the situation to a scale of 4, meaning that circumstances are "Somewhat responsible". This is inferred from the narrator's statement "I've recently heard of stories" which implies that the narrator is basing their opinion on what they have heard rather than on their own experience or knowledge. The narrator may be aware of anecdotal evidence of people getting sick after being vaccinated, but is unsure as to the extent that this is the case. Thus, the narrator believes that there is an element of luck or unpredictability involved in the situation that needs to be considered.

1

Figure 8: A comprehensive example from COVIDET-APPRAISALS (Page 1/8).



4. To what extent did the narrator think that they were able to COPE with the consequences of the event? (For example, if the narrator thought that they had the resources or the knowledge to make the situation better, or at least manageable.)

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** The narrator seems to be having a hard time coping with the idea of long-term illness. They mention that these recent updates to the long COVID situation make them want to 'stay inside forever like a hermit'.

🤖**ChatGPT**

- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator seems to be feeling overwhelmed and hopeless about the situation, indicating a low level of coping ability. They express a desire to stay inside forever like a hermit, which suggests a lack of confidence in their ability to manage the situation. The information they have learned has caused them to question whether the pandemic will ever truly end, further contributing to their sense of helplessness.

🤖**Alpaca**

- ➔ **[Scale]: 6. Narrator thought they were: Mostly able to cope.**
- ➔ **[Rationale]:** The narrator believed they had the knowledge and resources to make the situation manageable, but not necessarily completely able to cope with it. They realized that they would need to take precautions and reach out for help in order to make it through the difficult time.

5. To what extent did the narrator think that the situation was RELEVANT to their concerns and goals? (For example, if the narrator thought that the situation was personally important to what they desire.)

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 7**
- ➔ **[Rationale]:** The narrator seems to value their health and wellbeing. The situation is very relevant to their desire to stay healthy and take precautions against getting sick.

🤖**ChatGPT**

- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator expresses concern about the possibility of long covid even after being vaccinated, which directly affects their desire for freedom and ability to leave their home. Therefore, they likely consider the situation to be highly relevant to their concerns and goals.

🤖**Alpaca**

- ➔ **[Scale]: 9**
- ➔ **[Rationale]:** The narrator thought that the situation was completely relevant to their goals because it indicates the potential for a huge problem due to vaccinations not protecting against long covid, and they want to take precautions to prevent any kind of illness. They realize that the risk of getting sick is not going away and are trying to take steps to protect themselves.

6. To what extent did the narrator think that they needed to ATTEND to the situation further? (For example, if the narrator thought that the situation was either very complicated, dangerous, or interesting, that required them to pay more attention to deal with it.)

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 9**
- ➔ **[Rationale]:** The narrator seems to think the long-COVID situation is very complicated and dangerous. They see it as a long-term risk to their health and want to better understand how to protect themselves.

🤖**ChatGPT**

- ➔ **[Scale]: 7**
- ➔ **[Rationale]:** The narrator seems to be concerned about the possibility of long covid even after being vaccinated, which could potentially be a huge problem. They express a desire to stay inside forever like a hermit, indicating a significant level of worry. However, they do not provide any specific actions they plan to take or further information they need to gather, so the level of attention needed is not the highest possible.

🤖**Alpaca**

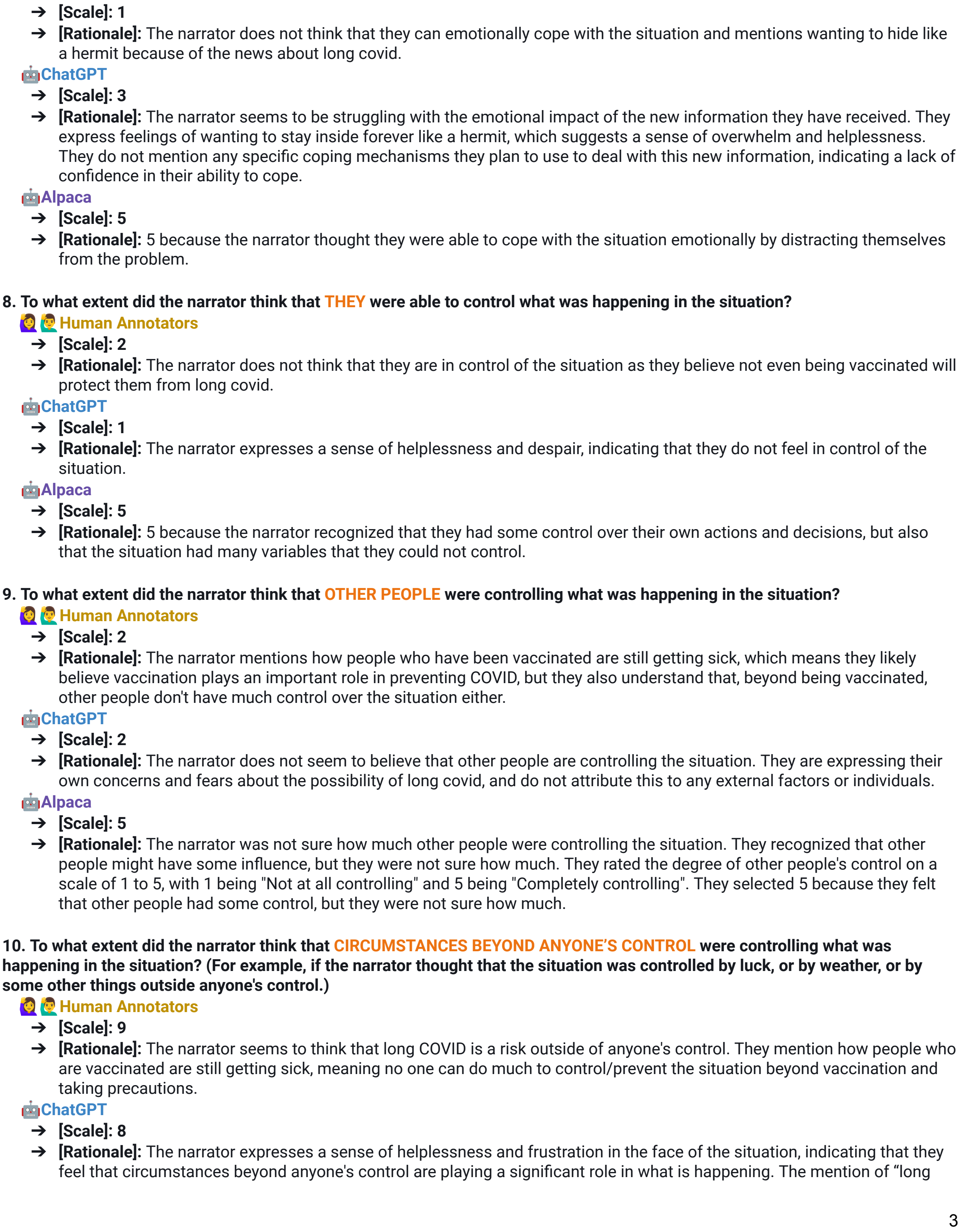
- ➔ **[Scale]: 9**
- ➔ **[Rationale]:** The narrator thought that the situation was very complicated and needed to be dealt with completely. They understood that the situation was dangerous and wanted to attend to it as soon as possible.

7. To what extent did the narrator think that they were able to EMOTIONALLY COPE with the consequences of the event? (For example, instead of dealing with the problem in the situation directly, the narrator thought that they are able to cope with the situation via other means such as distracting themselves from the problem by being busy, eating comfort food or drinking alcohol.)

🙋‍♀️🙋‍♂️**Human Annotators**

2

Figure 9: A comprehensive example from CovidET-Appraisals (Page 2/8).



- ➔ **[Scale]:** 1
- ➔ **[Rationale]:** The narrator does not think that they can emotionally cope with the situation and mentions wanting to hide like a hermit because of the news about long covid.

🤖**ChatGPT**
- ➔ **[Scale]:** 3
- ➔ **[Rationale]:** The narrator seems to be struggling with the emotional impact of the new information they have received. They express feelings of wanting to stay inside forever like a hermit, which suggests a sense of overwhelm and helplessness. They do not mention any specific coping mechanisms they plan to use to deal with this new information, indicating a lack of confidence in their ability to cope.

🤖**Alpaca**
- ➔ **[Scale]:** 5
- ➔ **[Rationale]:** 5 because the narrator thought they were able to cope with the situation emotionally by distracting themselves from the problem.

8. To what extent did the narrator think that THEY were able to control what was happening in the situation?

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]:** 2
- ➔ **[Rationale]:** The narrator does not think that they are in control of the situation as they believe not even being vaccinated will protect them from long covid.

🤖**ChatGPT**
- ➔ **[Scale]:** 1
- ➔ **[Rationale]:** The narrator expresses a sense of helplessness and despair, indicating that they do not feel in control of the situation.

🤖**Alpaca**
- ➔ **[Scale]:** 5
- ➔ **[Rationale]:** 5 because the narrator recognized that they had some control over their own actions and decisions, but also that the situation had many variables that they could not control.

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

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]:** 2
- ➔ **[Rationale]:** The narrator mentions how people who have been vaccinated are still getting sick, which means they likely believe vaccination plays an important role in preventing COVID, but they also understand that, beyond being vaccinated, other people don't have much control over the situation either.

🤖**ChatGPT**
- ➔ **[Scale]:** 2
- ➔ **[Rationale]:** The narrator does not seem to believe that other people are controlling the situation. They are expressing their own concerns and fears about the possibility of long covid, and do not attribute this to any external factors or individuals.

🤖**Alpaca**
- ➔ **[Scale]:** 5
- ➔ **[Rationale]:** The narrator was not sure how much other people were controlling the situation. They recognized that other people might have some influence, but they were not sure how much. They rated the degree of other people's control on a scale of 1 to 5, with 1 being "Not at all controlling" and 5 being "Completely controlling". They selected 5 because they felt that other people had some control, but they were not sure how much.

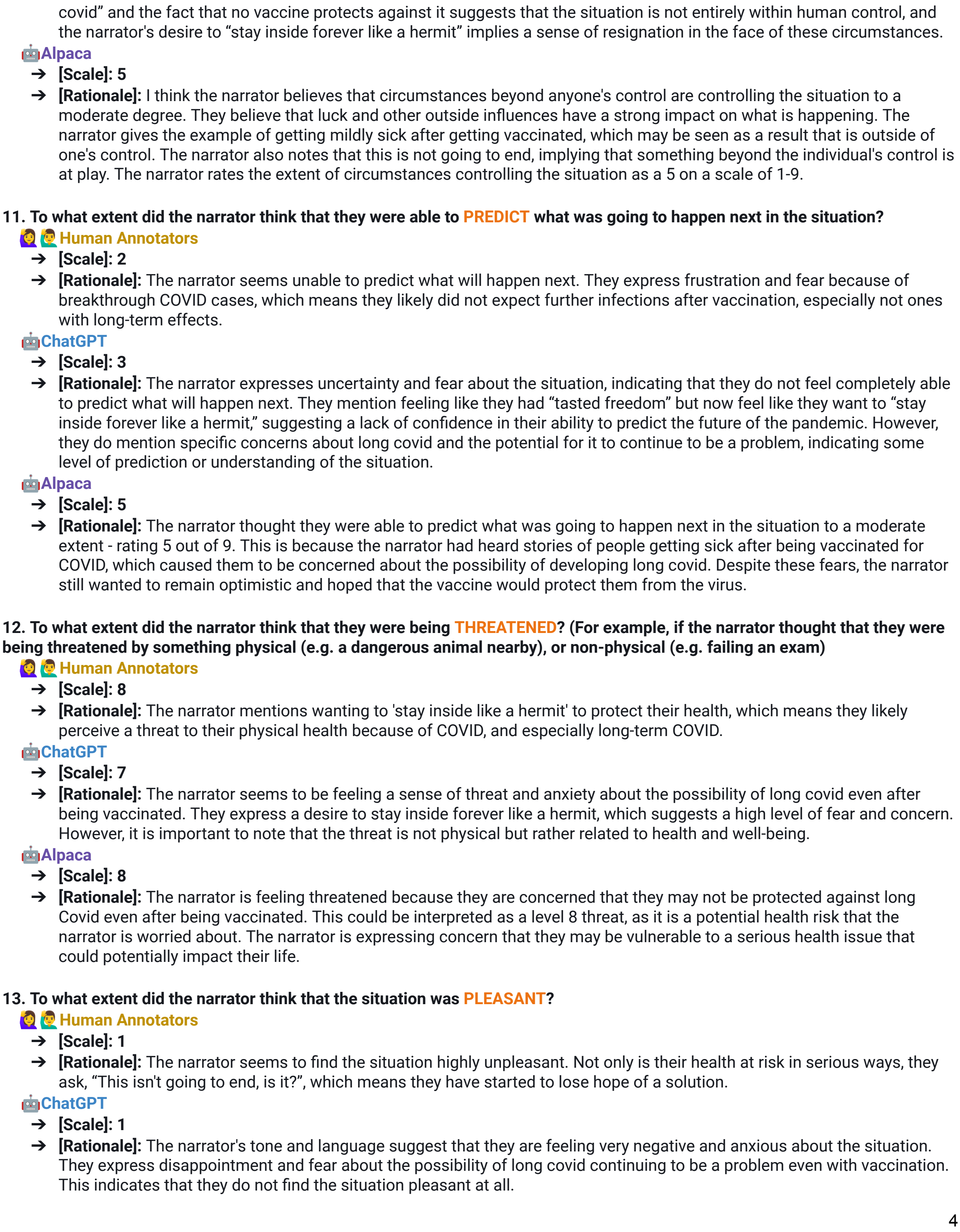
10. To what extent did the narrator think that CIRCUMSTANCES BEYOND ANYONE'S CONTROL were controlling what was happening in the situation? (For example, if the narrator thought that the situation was controlled by luck, or by weather, or by some other things outside anyone's control.)

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]:** 9
- ➔ **[Rationale]:** The narrator seems to think that long COVID is a risk outside of anyone's control. They mention how people who are vaccinated are still getting sick, meaning no one can do much to control/prevent the situation beyond vaccination and taking precautions.

🤖**ChatGPT**
- ➔ **[Scale]:** 8
- ➔ **[Rationale]:** The narrator expresses a sense of helplessness and frustration in the face of the situation, indicating that they feel that circumstances beyond anyone's control are playing a significant role in what is happening. The mention of "long

3

Figure 10: A comprehensive example from CovidET-Appraisals (Page 3/8).



covid" and the fact that no vaccine protects against it suggests that the situation is not entirely within human control, and the narrator's desire to "stay inside forever like a hermit" implies a sense of resignation in the face of these circumstances.

🤖**Alpaca**
➔ **[Scale]: 5**
➔ **[Rationale]:** I think the narrator believes that circumstances beyond anyone's control are controlling the situation to a moderate degree. They believe that luck and other outside influences have a strong impact on what is happening. The narrator gives the example of getting mildly sick after getting vaccinated, which may be seen as a result that is outside of one's control. The narrator also notes that this is not going to end, implying that something beyond the individual's control is at play. The narrator rates the extent of circumstances controlling the situation as a 5 on a scale of 1-9.

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

🙋‍♀️🙋‍♂️**Human Annotators**
➔ **[Scale]: 2**
➔ **[Rationale]:** The narrator seems unable to predict what will happen next. They express frustration and fear because of breakthrough COVID cases, which means they likely did not expect further infections after vaccination, especially not ones with long-term effects.

🤖**ChatGPT**
➔ **[Scale]: 3**
➔ **[Rationale]:** The narrator expresses uncertainty and fear about the situation, indicating that they do not feel completely able to predict what will happen next. They mention feeling like they had "tasted freedom" but now feel like they want to "stay inside forever like a hermit," suggesting a lack of confidence in their ability to predict the future of the pandemic. However, they do mention specific concerns about long covid and the potential for it to continue to be a problem, indicating some level of prediction or understanding of the situation.

🤖**Alpaca**
➔ **[Scale]: 5**
➔ **[Rationale]:** The narrator thought they were able to predict what was going to happen next in the situation to a moderate extent - rating 5 out of 9. This is because the narrator had heard stories of people getting sick after being vaccinated for COVID, which caused them to be concerned about the possibility of developing long covid. Despite these fears, the narrator still wanted to remain optimistic and hoped that the vaccine would protect them from the virus.

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)

🙋‍♀️🙋‍♂️**Human Annotators**
➔ **[Scale]: 8**
➔ **[Rationale]:** The narrator mentions wanting to 'stay inside like a hermit' to protect their health, which means they likely perceive a threat to their physical health because of COVID, and especially long-term COVID.

🤖**ChatGPT**
➔ **[Scale]: 7**
➔ **[Rationale]:** The narrator seems to be feeling a sense of threat and anxiety about the possibility of long covid even after being vaccinated. They express a desire to stay inside forever like a hermit, which suggests a high level of fear and concern. However, it is important to note that the threat is not physical but rather related to health and well-being.

🤖**Alpaca**
➔ **[Scale]: 8**
➔ **[Rationale]:** The narrator is feeling threatened because they are concerned that they may not be protected against long Covid even after being vaccinated. This could be interpreted as a level 8 threat, as it is a potential health risk that the narrator is worried about. The narrator is expressing concern that they may be vulnerable to a serious health issue that could potentially impact their life.

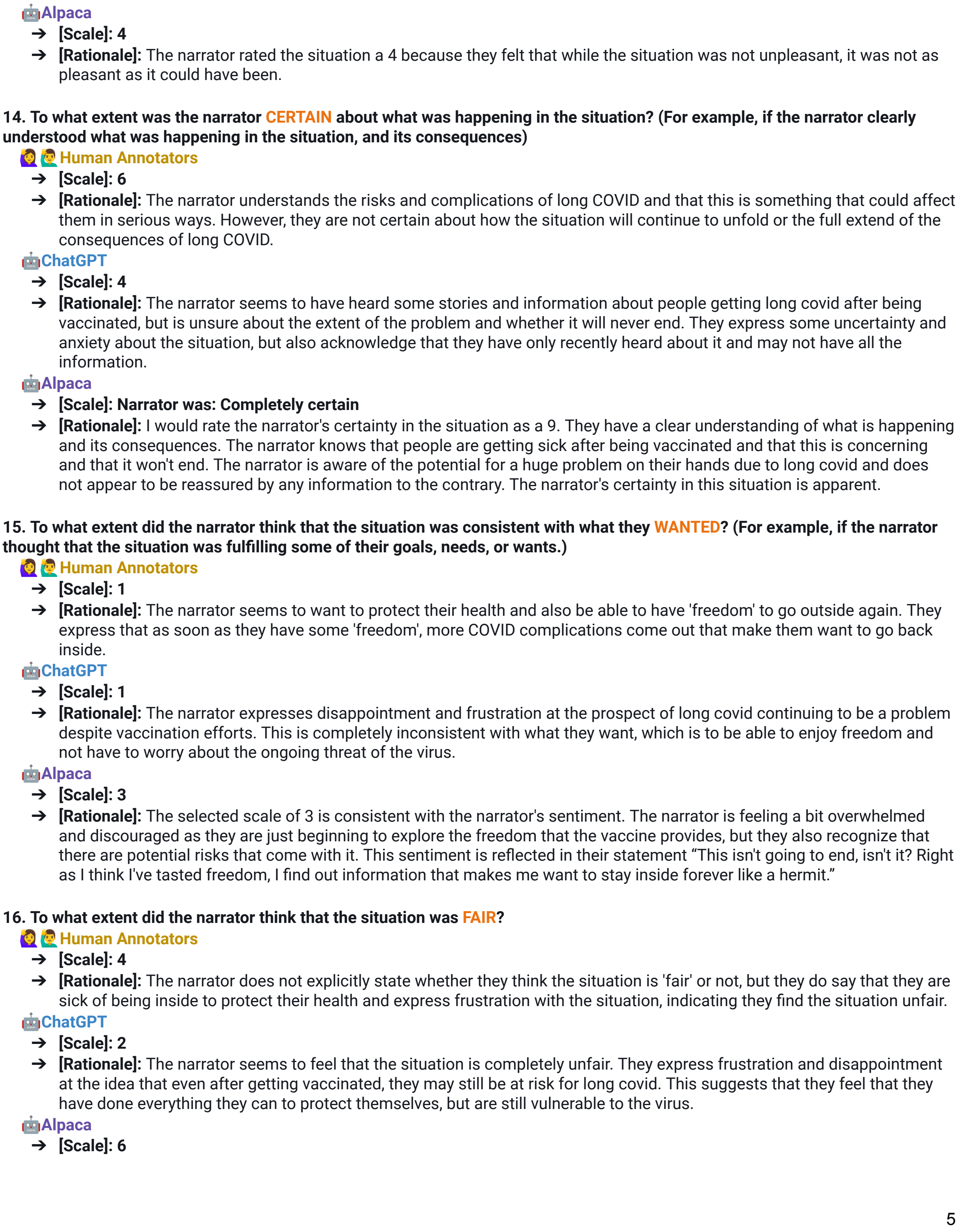
13. To what extent did the narrator think that the situation was PLEASANT?

🙋‍♀️🙋‍♂️**Human Annotators**
➔ **[Scale]: 1**
➔ **[Rationale]:** The narrator seems to find the situation highly unpleasant. Not only is their health at risk in serious ways, they ask, "This isn't going to end, is it?", which means they have started to lose hope of a solution.

🤖**ChatGPT**
➔ **[Scale]: 1**
➔ **[Rationale]:** The narrator's tone and language suggest that they are feeling very negative and anxious about the situation. They express disappointment and fear about the possibility of long covid continuing to be a problem even with vaccination. This indicates that they do not find the situation pleasant at all.

4

Figure 11: A comprehensive example from COVIDET-APPRAISALS (Page 4/8).



🤖**Alpaca**

- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** The narrator rated the situation a 4 because they felt that while the situation was not unpleasant, it was not as pleasant as it could have been.

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)

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 6**
- ➔ **[Rationale]:** The narrator understands the risks and complications of long COVID and that this is something that could affect them in serious ways. However, they are not certain about how the situation will continue to unfold or the full extend of the consequences of long COVID.

🤖**ChatGPT**

- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** The narrator seems to have heard some stories and information about people getting long covid after being vaccinated, but is unsure about the extent of the problem and whether it will never end. They express some uncertainty and anxiety about the situation, but also acknowledge that they have only recently heard about it and may not have all the information.

🤖**Alpaca**

- ➔ **[Scale]: Narrator was: Completely certain**
- ➔ **[Rationale]:** I would rate the narrator's certainty in the situation as a 9. They have a clear understanding of what is happening and its consequences. The narrator knows that people are getting sick after being vaccinated and that this is concerning and that it won't end. The narrator is aware of the potential for a huge problem on their hands due to long covid and does not appear to be reassured by any information to the contrary. The narrator's certainty in this situation is apparent.

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.)

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** The narrator seems to want to protect their health and also be able to have 'freedom' to go outside again. They express that as soon as they have some 'freedom', more COVID complications come out that make them want to go back inside.

🤖**ChatGPT**

- ➔ **[Scale]: 1**
- ➔ **[Rationale]:** The narrator expresses disappointment and frustration at the prospect of long covid continuing to be a problem despite vaccination efforts. This is completely inconsistent with what they want, which is to be able to enjoy freedom and not have to worry about the ongoing threat of the virus.

🤖**Alpaca**

- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The selected scale of 3 is consistent with the narrator's sentiment. The narrator is feeling a bit overwhelmed and discouraged as they are just beginning to explore the freedom that the vaccine provides, but they also recognize that there are potential risks that come with it. This sentiment is reflected in their statement "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."

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

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** The narrator does not explicitly state whether they think the situation is 'fair' or not, but they do say that they are sick of being inside to protect their health and express frustration with the situation, indicating they find the situation unfair.

🤖**ChatGPT**

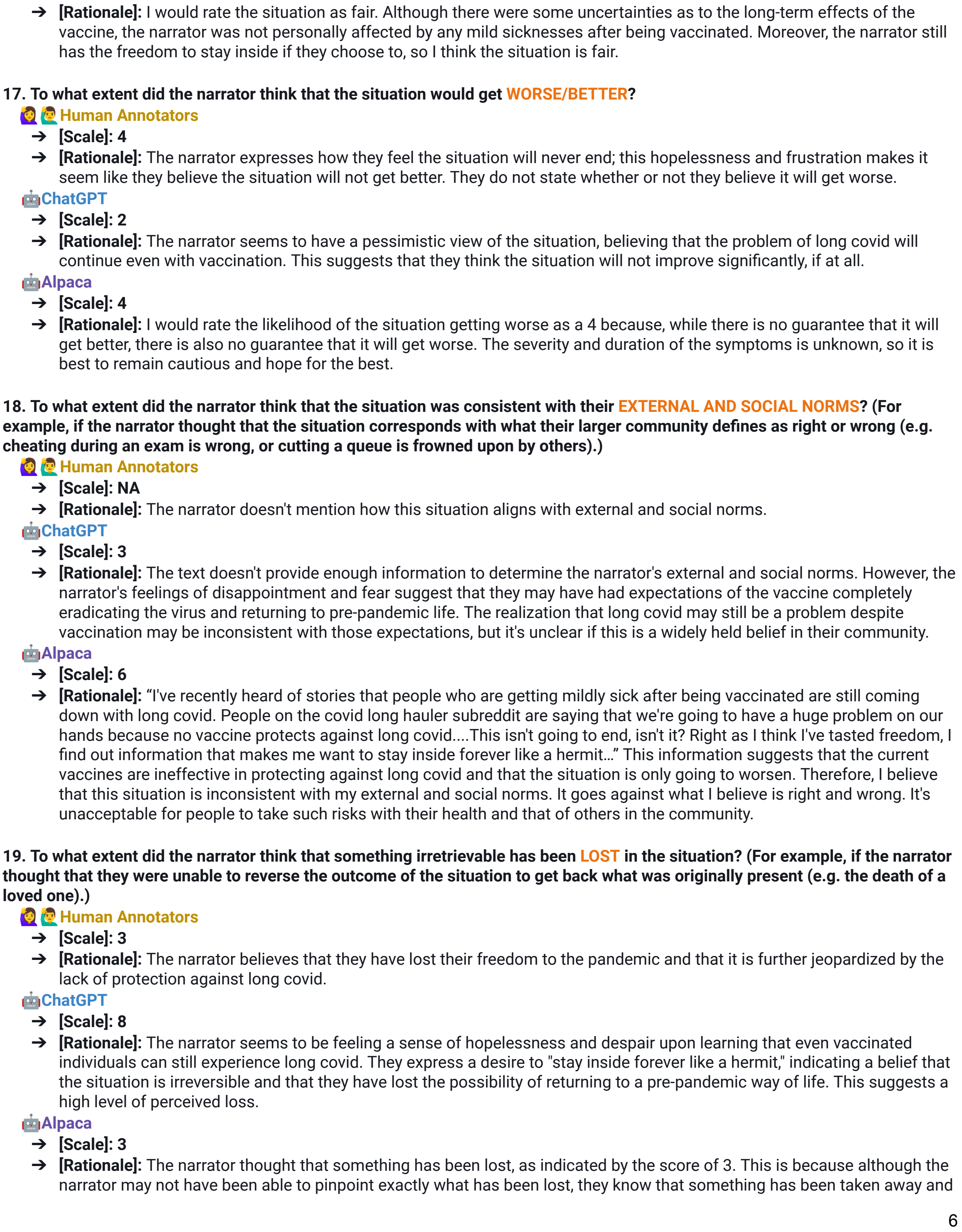
- ➔ **[Scale]: 2**
- ➔ **[Rationale]:** The narrator seems to feel that the situation is completely unfair. They express frustration and disappointment at the idea that even after getting vaccinated, they may still be at risk for long covid. This suggests that they feel that they have done everything they can to protect themselves, but are still vulnerable to the virus.

🤖**Alpaca**

- ➔ **[Scale]: 6**

5

Figure 12: A comprehensive example from COVIDET-APPRAISALS (Page 5/8).



- ➔ **[Rationale]:** I would rate the situation as fair. Although there were some uncertainties as to the long-term effects of the vaccine, the narrator was not personally affected by any mild sicknesses after being vaccinated. Moreover, the narrator still has the freedom to stay inside if they choose to, so I think the situation is fair.

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

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** The narrator expresses how they feel the situation will never end; this hopelessness and frustration makes it seem like they believe the situation will not get better. They do not state whether or not they believe it will get worse.

🤖**ChatGPT**
- ➔ **[Scale]: 2**
- ➔ **[Rationale]:** The narrator seems to have a pessimistic view of the situation, believing that the problem of long covid will continue even with vaccination. This suggests that they think the situation will not improve significantly, if at all.

🤖**Alpaca**
- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** I would rate the likelihood of the situation getting worse as a 4 because, while there is no guarantee that it will get better, there is also no guarantee that it will get worse. The severity and duration of the symptoms is unknown, so it is best to remain cautious and hope for the best.

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).)

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: NA**
- ➔ **[Rationale]:** The narrator doesn't mention how this situation aligns with external and social norms.

🤖**ChatGPT**
- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The text doesn't provide enough information to determine the narrator's external and social norms. However, the narrator's feelings of disappointment and fear suggest that they may have had expectations of the vaccine completely eradicating the virus and returning to pre-pandemic life. The realization that long covid may still be a problem despite vaccination may be inconsistent with those expectations, but it's unclear if this is a widely held belief in their community.

🤖**Alpaca**
- ➔ **[Scale]: 6**
- ➔ **[Rationale]:** "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…" This information suggests that the current vaccines are ineffective in protecting against long covid and that the situation is only going to worsen. Therefore, I believe that this situation is inconsistent with my external and social norms. It goes against what I believe is right and wrong. It's unacceptable for people to take such risks with their health and that of others in the community.

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).)

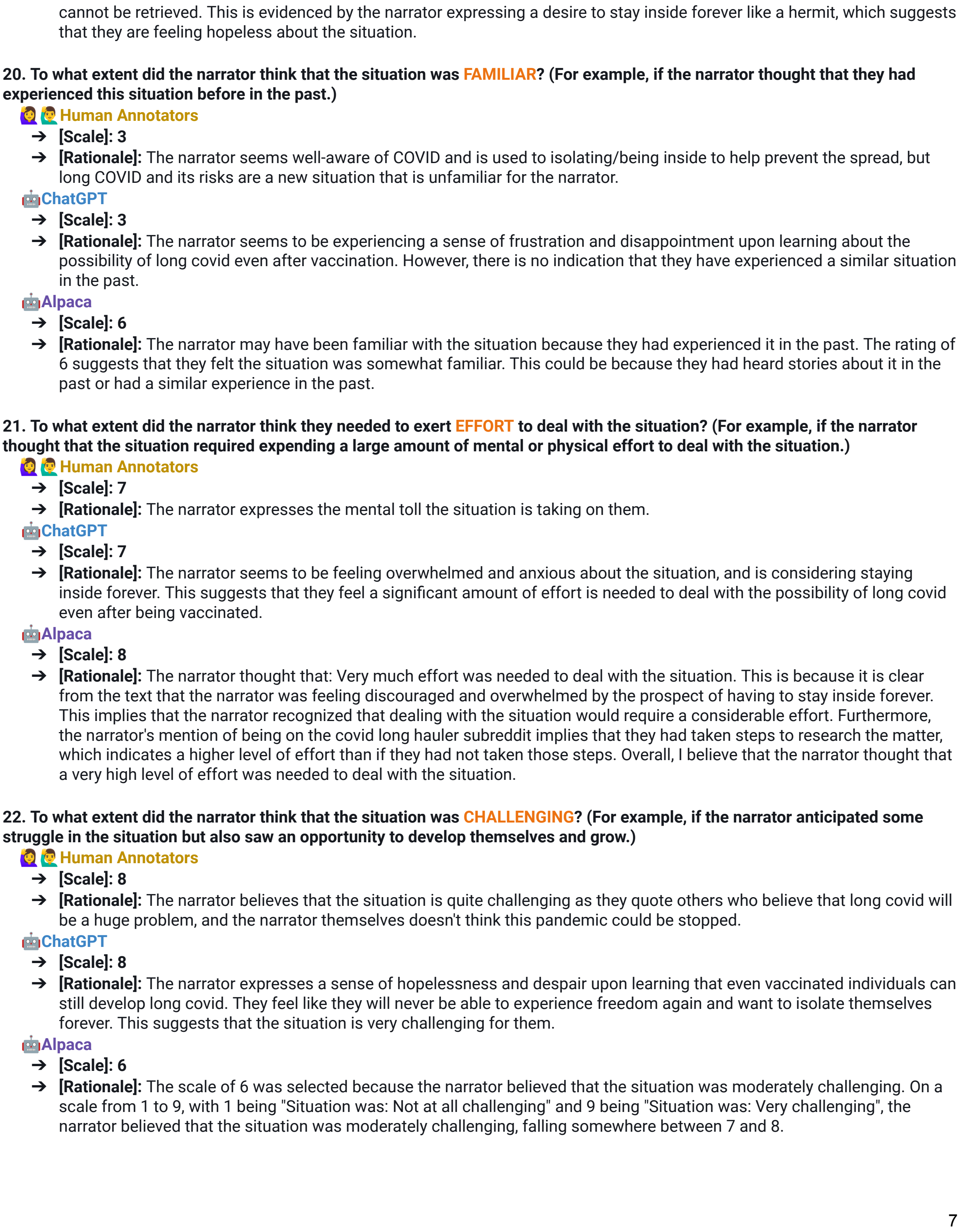
🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator believes that they have lost their freedom to the pandemic and that it is further jeopardized by the lack of protection against long covid.

🤖**ChatGPT**
- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator seems to be feeling a sense of hopelessness and despair upon learning that even vaccinated individuals can still experience long covid. They express a desire to "stay inside forever like a hermit," indicating a belief that the situation is irreversible and that they have lost the possibility of returning to a pre-pandemic way of life. This suggests a high level of perceived loss.

🤖**Alpaca**
- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator thought that something has been lost, as indicated by the score of 3. This is because although the narrator may not have been able to pinpoint exactly what has been lost, they know that something has been taken away and

6

Figure 13: A comprehensive example from COVIDET-APPRAISALS (Page 6/8).



cannot be retrieved. This is evidenced by the narrator expressing a desire to stay inside forever like a hermit, which suggests that they are feeling hopeless about the situation.

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.)

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator seems well-aware of COVID and is used to isolating/being inside to help prevent the spread, but long COVID and its risks are a new situation that is unfamiliar for the narrator.

🤖**ChatGPT**
- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator seems to be experiencing a sense of frustration and disappointment upon learning about the possibility of long covid even after vaccination. However, there is no indication that they have experienced a similar situation in the past.

🤖**Alpaca**
- ➔ **[Scale]: 6**
- ➔ **[Rationale]:** The narrator may have been familiar with the situation because they had experienced it in the past. The rating of 6 suggests that they felt the situation was somewhat familiar. This could be because they had heard stories about it in the past or had a similar experience in the past.

21. To what extent did the narrator think they needed to exert EFFORT to deal with the situation? (For example, if the narrator thought that the situation required expending a large amount of mental or physical effort to deal with the situation.)

🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 7**
- ➔ **[Rationale]:** The narrator expresses the mental toll the situation is taking on them.

🤖**ChatGPT**
- ➔ **[Scale]: 7**
- ➔ **[Rationale]:** The narrator seems to be feeling overwhelmed and anxious about the situation, and is considering staying inside forever. This suggests that they feel a significant amount of effort is needed to deal with the possibility of long covid even after being vaccinated.

🤖**Alpaca**
- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator thought that: Very much effort was needed to deal with the situation. This is because it is clear from the text that the narrator was feeling discouraged and overwhelmed by the prospect of having to stay inside forever. This implies that the narrator recognized that dealing with the situation would require a considerable effort. Furthermore, the narrator's mention of being on the covid long hauler subreddit implies that they had taken steps to research the matter, which indicates a higher level of effort than if they had not taken those steps. Overall, I believe that the narrator thought that a very high level of effort was needed to deal with the situation.

22. To what extent did the narrator think that the situation was CHALLENGING? (For example, if the narrator anticipated some struggle in the situation but also saw an opportunity to develop themselves and grow.)

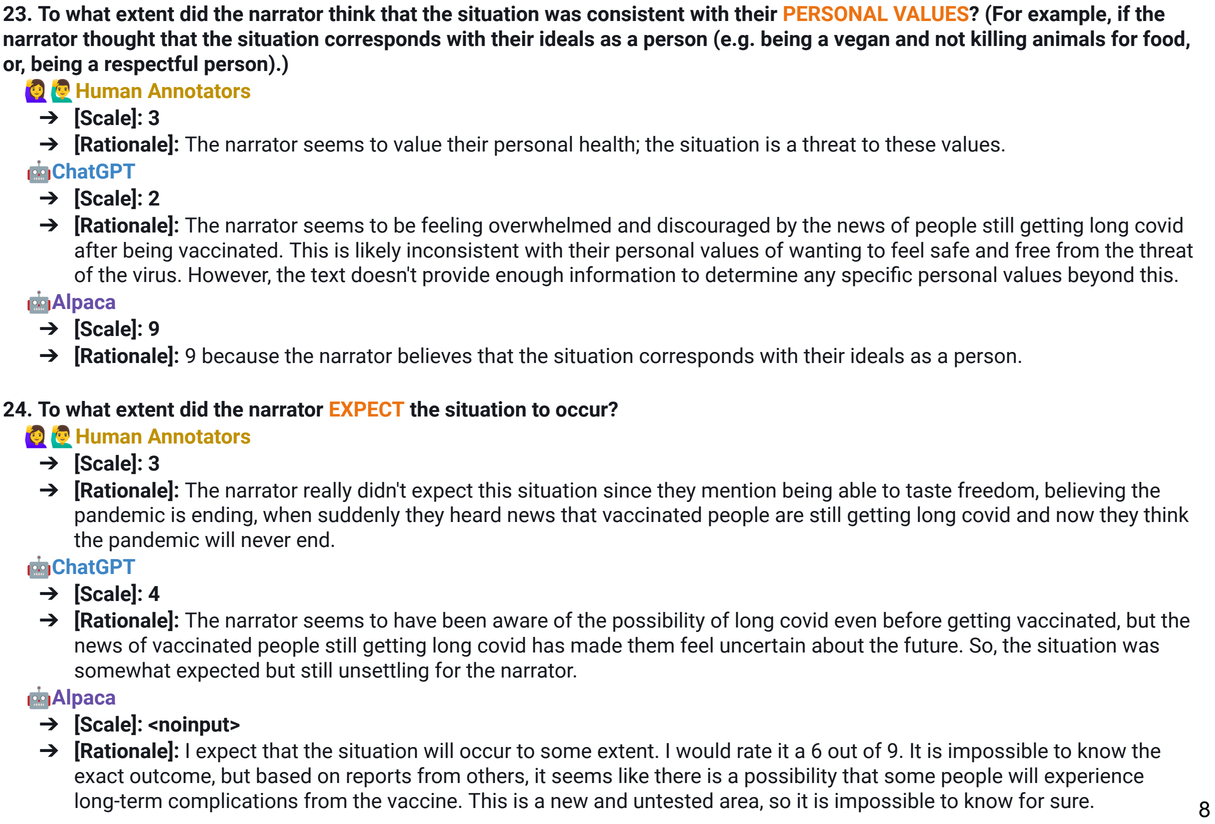
🙋‍♀️🙋‍♂️**Human Annotators**
- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator believes that the situation is quite challenging as they quote others who believe that long covid will be a huge problem, and the narrator themselves doesn't think this pandemic could be stopped.

🤖**ChatGPT**
- ➔ **[Scale]: 8**
- ➔ **[Rationale]:** The narrator expresses a sense of hopelessness and despair upon learning that even vaccinated individuals can still develop long covid. They feel like they will never be able to experience freedom again and want to isolate themselves forever. This suggests that the situation is very challenging for them.

🤖**Alpaca**
- ➔ **[Scale]: 6**
- ➔ **[Rationale]:** The scale of 6 was selected because the narrator believed that the situation was moderately challenging. On a scale from 1 to 9, with 1 being "Situation was: Not at all challenging" and 9 being "Situation was: Very challenging", the narrator believed that the situation was moderately challenging, falling somewhere between 7 and 8.

7

Figure 14: A comprehensive example from COVIDET-APPRAISALS (Page 7/8).



23. To what extent did the narrator think that the situation was consistent with their PERSONAL VALUES? (For example, if the narrator thought that the situation corresponds with their ideals as a person (e.g. being a vegan and not killing animals for food, or, being a respectful person).)

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator seems to value their personal health; the situation is a threat to these values.

🤖**ChatGPT**

- ➔ **[Scale]: 2**
- ➔ **[Rationale]:** The narrator seems to be feeling overwhelmed and discouraged by the news of people still getting long covid after being vaccinated. This is likely inconsistent with their personal values of wanting to feel safe and free from the threat of the virus. However, the text doesn't provide enough information to determine any specific personal values beyond this.

🤖**Alpaca**

- ➔ **[Scale]: 9**
- ➔ **[Rationale]:** 9 because the narrator believes that the situation corresponds with their ideals as a person.

24. To what extent did the narrator EXPECT the situation to occur?

🙋‍♀️🙋‍♂️**Human Annotators**

- ➔ **[Scale]: 3**
- ➔ **[Rationale]:** The narrator really didn't expect 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.

🤖**ChatGPT**

- ➔ **[Scale]: 4**
- ➔ **[Rationale]:** The narrator seems to have been aware of the possibility of long covid even before getting vaccinated, but the news of vaccinated people still getting long covid has made them feel uncertain about the future. So, the situation was somewhat expected but still unsettling for the narrator.

🤖**Alpaca**

- ➔ **[Scale]: <noinput>**
- ➔ **[Rationale]:** I expect that the situation will occur to some extent. I would rate it a 6 out of 9. It is impossible to know the exact outcome, but based on reports from others, it seems like there is a possibility that some people will experience long-term complications from the vaccine. This is a new and untested area, so it is impossible to know for sure.

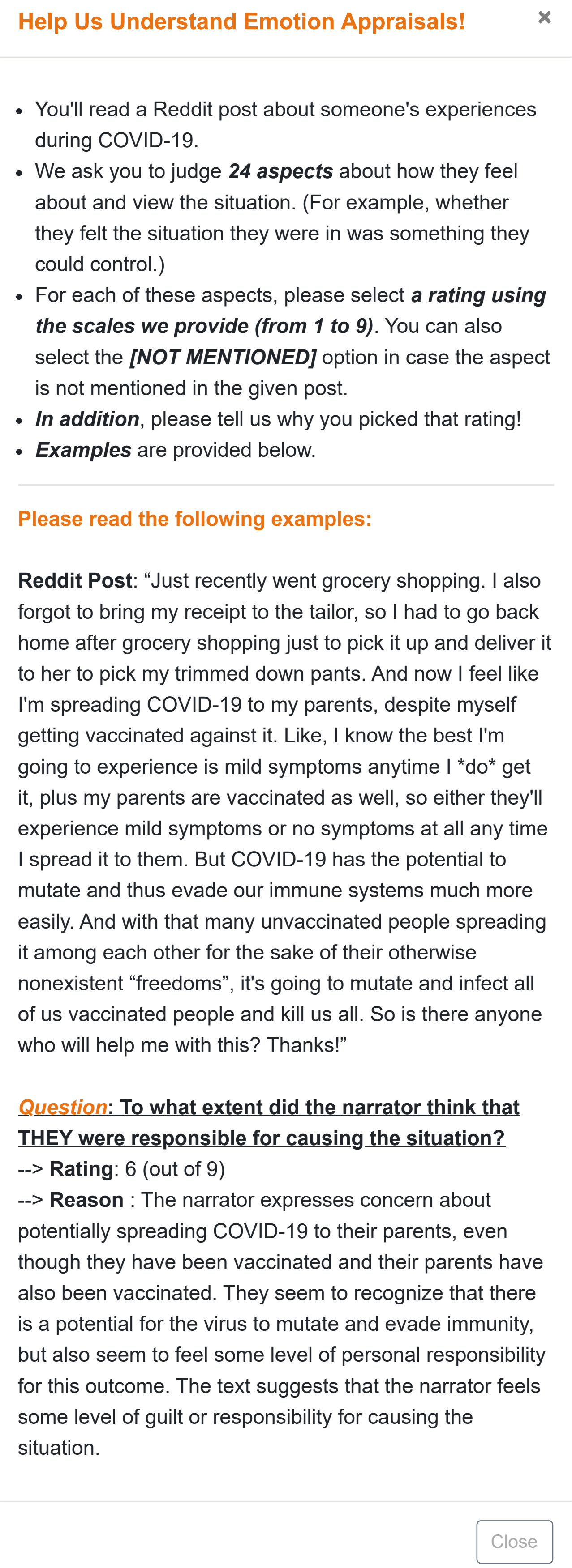
Figure 15: A comprehensive example from COVIDET-APPRAISALS (Page 8/8).

	BLEU			ROUGE			BERTSCORE	
	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	RE-SCALED
EXPERTS	0.161	0.085	0.046	0.336	0.120	0.262	0.895	0.379
BASELINE-P	0.174	0.108	0.076	0.336	0.142	0.279	0.893	0.364
BASELINE-D	0.152	0.087	0.059	0.305	0.118	0.247	0.888	0.337

Table 4: The full rationale agreement statistics. BASELINE-P denotes "baseline (same dimension, *different posts*)", and BASELINE-D denotes "baseline (same post, *different dimensions*)".

	BLEU			ROUGE			BERTSCORE	
	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	RE-SCALED
CHATGPT								
ALPACA-7B								
ALPACA-13B								
DOLLY-7B								
DOLLY-12B								
FLAN-T5-XXL								

Table 5: The full rationale statistics measured for LLMs' responses against the gold annotations.



Help Us Understand Emotion Appraisals!

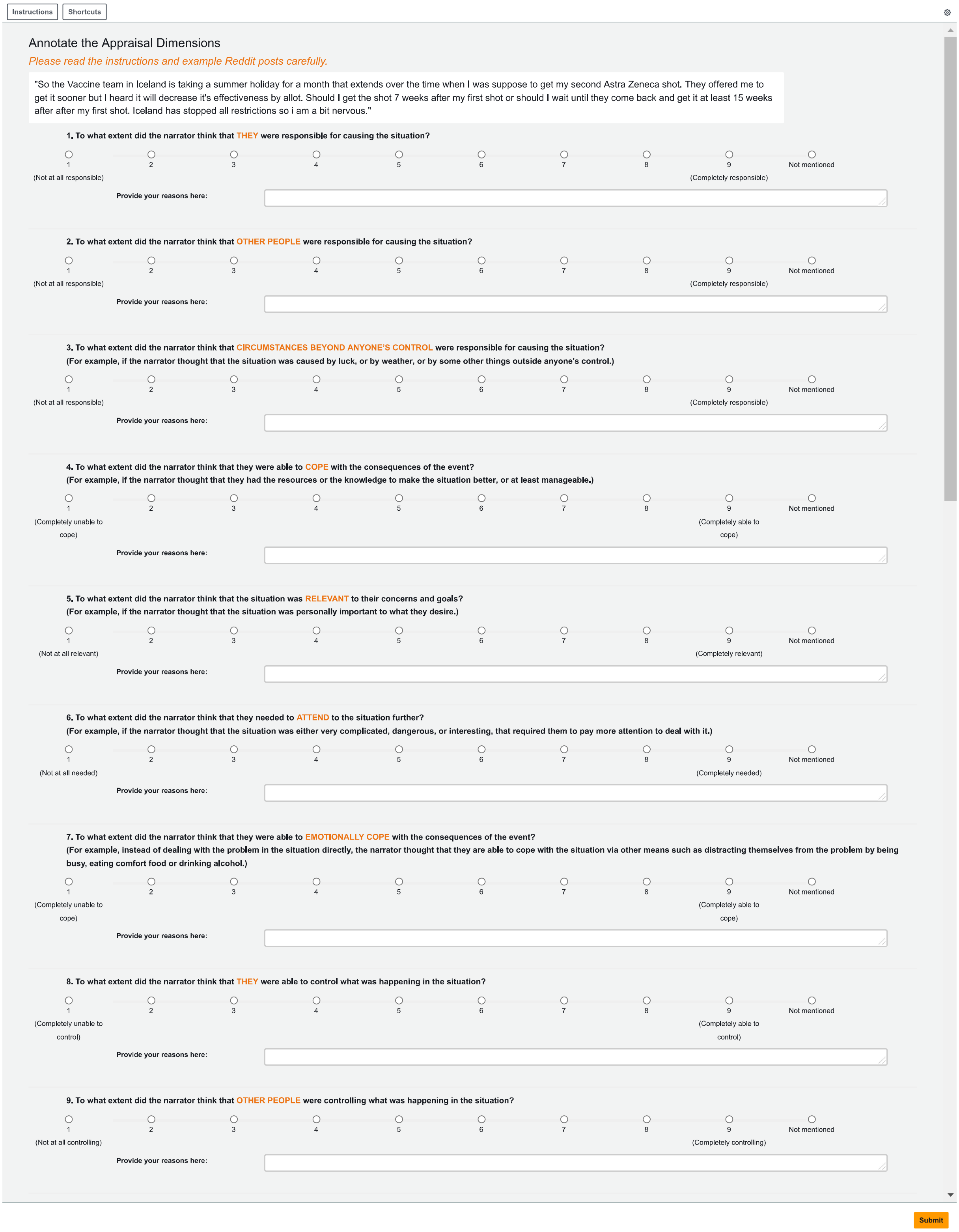
- You'll read a Reddit post about someone's experiences during COVID-19.
- We ask you to judge ***24 aspects*** about how they feel about and view the situation. (For example, whether they felt the situation they were in was something they could control.)
- For each of these aspects, please select ***a rating using the scales we provide (from 1 to 9)***. You can also select the ***[NOT MENTIONED]*** option in case the aspect is not mentioned in the given post.
- ***In addition***, please tell us why you picked that rating!
- ***Examples*** are provided below.

Please read the following examples:

Reddit Post: “Just recently went grocery shopping. I also forgot to bring my receipt to the tailor, so I had to go back home after grocery shopping just to pick it up and deliver it to her to pick my trimmed down pants. And now I feel like I'm spreading COVID-19 to my parents, despite myself getting vaccinated against it. Like, I know the best I'm going to experience is mild symptoms anytime I *do* get it, plus my parents are vaccinated as well, so either they'll experience mild symptoms or no symptoms at all any time I spread it to them. But COVID-19 has the potential to mutate and thus evade our immune systems much more easily. And with that many unvaccinated people spreading it among each other for the sake of their otherwise nonexistent “freedoms”, it's going to mutate and infect all of us vaccinated people and kill us all. So is there anyone who will help me with this? Thanks!”

***Question*: To what extent did the narrator think that THEY were responsible for causing the situation?**
--> **Rating**: 6 (out of 9)
--> **Reason** : The narrator expresses concern about potentially spreading COVID-19 to their parents, even though they have been vaccinated and their parents have also been vaccinated. They seem to recognize that there is a potential for the virus to mutate and evade immunity, but also seem to feel some level of personal responsibility for this outcome. The text suggests that the narrator feels some level of guilt or responsibility for causing the situation.

Figure 16: Instructions to annotators for COVIDET-APPRAISALS.



Annotate the Appraisal Dimensions

Please read the instructions and example Reddit posts carefully.

"So the Vaccine team in Iceland is taking a summer holiday for a month that extends over the time when I was suppose to get my second Astra Zeneca shot. They offered me to get it sooner but I heard it will decrease it's effectiveness by allot. Should I get the shot 7 weeks after my first shot or should I wait until they come back and get it at least 15 weeks after after my first shot. Iceland has stopped all restrictions so i am a bit nervous."

1. To what extent did the narrator think that THEY were responsible for causing the situation?

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

Provide your reasons here:

2. To what extent did the narrator think that OTHER PEOPLE were responsible for causing the situation?

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

Provide your reasons here:

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.)

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

Provide your reasons here:

4. To what extent did the narrator think that they were able to COPE with the consequences of the event?
(For example, if the narrator thought that they had the resources or the knowledge to make the situation better, or at least manageable.)

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

Provide your reasons here:

5. To what extent did the narrator think that the situation was RELEVANT to their concerns and goals?
(For example, if the narrator thought that the situation was personally important to what they desire.)

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

Provide your reasons here:

6. To what extent did the narrator think that they needed to ATTEND to the situation further?
(For example, if the narrator thought that the situation was either very complicated, dangerous, or interesting, that required them to pay more attention to deal with it.)

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

Provide your reasons here:

7. To what extent did the narrator think that they were able to EMOTIONALLY COPE with the consequences of the event?
(For example, instead of dealing with the problem in the situation directly, the narrator thought that they are able to cope with the situation via other means such as distracting themselves from the problem by being busy, eating comfort food or drinking alcohol.)

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

Provide your reasons here:

8. To what extent did the narrator think that THEY were able to control what was happening in the situation?

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

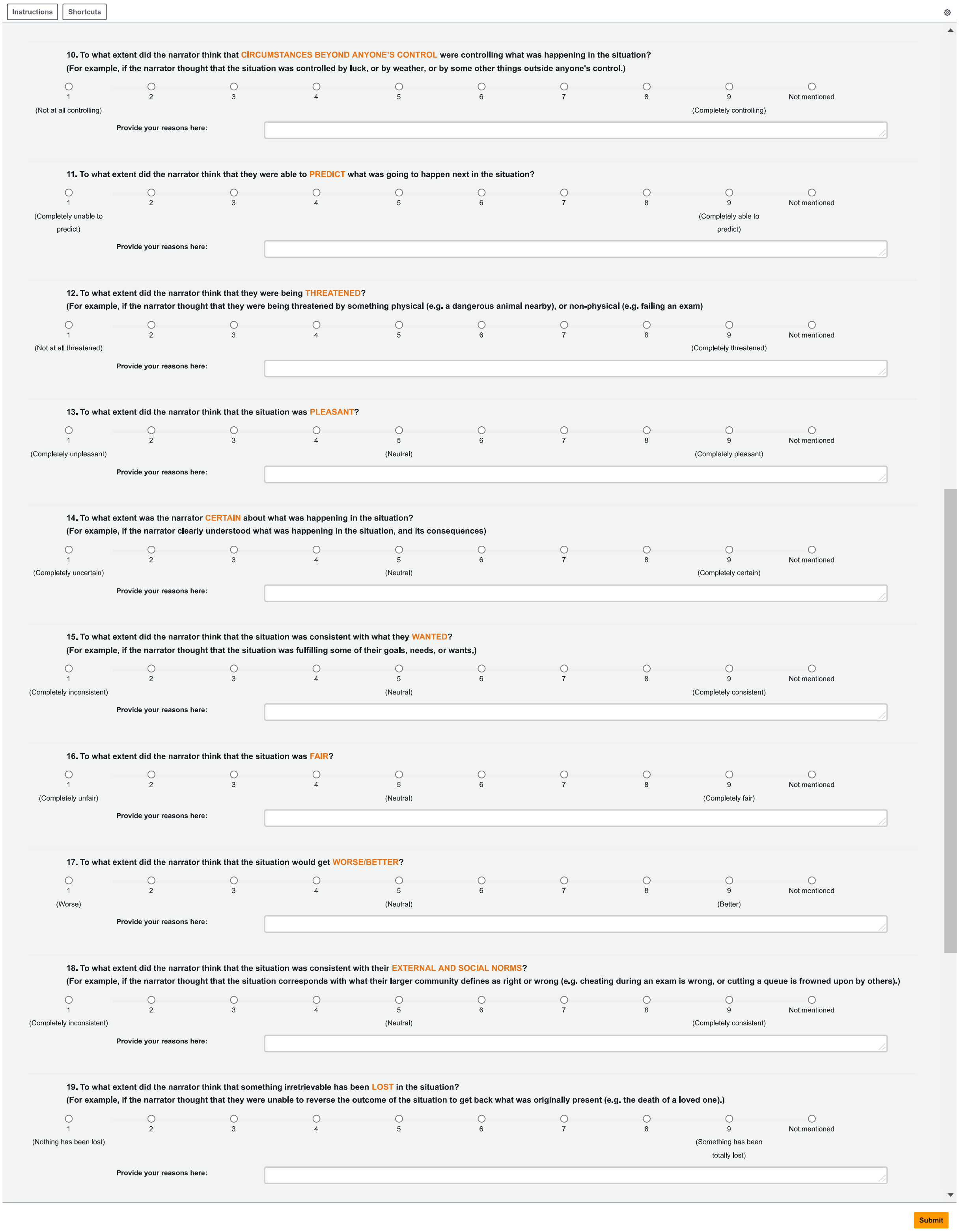
Provide your reasons here:

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

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

Provide your reasons here:

Figure 17: Annotation task layout for COVIDET-APPRAISALS (Page 1/3).



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

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

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

Provide your reasons here:

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

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

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

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

Provide your reasons here:

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

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

Provide your reasons here:

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

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

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

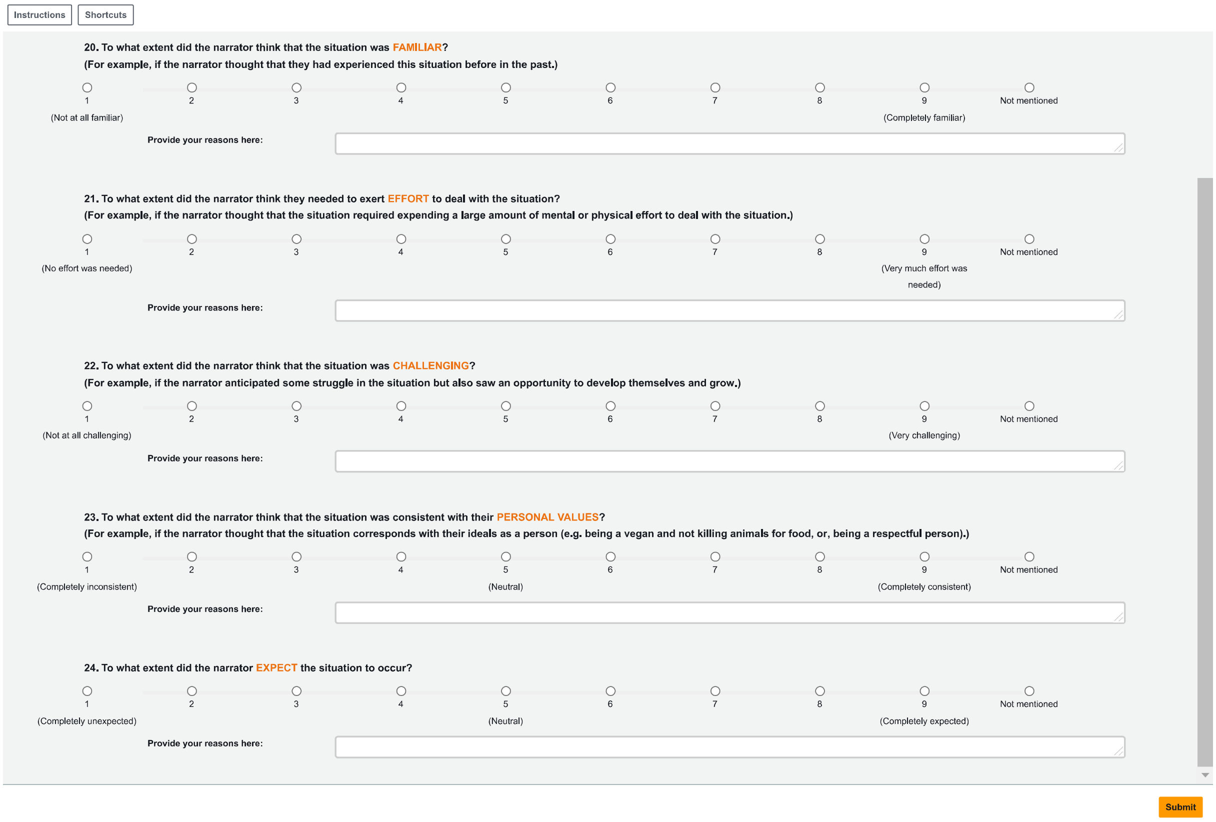
Provide your reasons here:

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

Provide your reasons here:

Figure 18: Annotation task layout for COVIDET-APPRAISALS (Page 2/3).



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.)

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

Provide your reasons here:

21. To what extent did the narrator think they needed to exert EFFORT to deal with the situation?
(For example, if the narrator thought that the situation required expending a large amount of mental or physical effort to deal with the situation.)

○ 1 (No effort was needed) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 (Very much effort was needed) ○ Not mentioned

Provide your reasons here:

22. To what extent did the narrator think that the situation was CHALLENGING?
(For example, if the narrator anticipated some struggle in the situation but also saw an opportunity to develop themselves and grow.)

○ 1 (Not at all challenging) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 (Very challenging) ○ Not mentioned

Provide your reasons here:

23. To what extent did the narrator think that the situation was consistent with their PERSONAL VALUES?
(For example, if the narrator thought that the situation corresponds with their ideals as a person (e.g. being a vegan and not killing animals for food, or, being a respectful person).)

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

Provide your reasons here:

24. To what extent did the narrator EXPECT the situation to occur?

○ 1 (Completely unexpected) ○ 2 ○ 3 ○ 4 ○ 5 (Neutral) ○ 6 ○ 7 ○ 8 ○ 9 (Completely expected) ○ Not mentioned

Provide your reasons here:

Figure 19: Annotation task layout for COVIDET-APPRAISALS (Page 3/3).

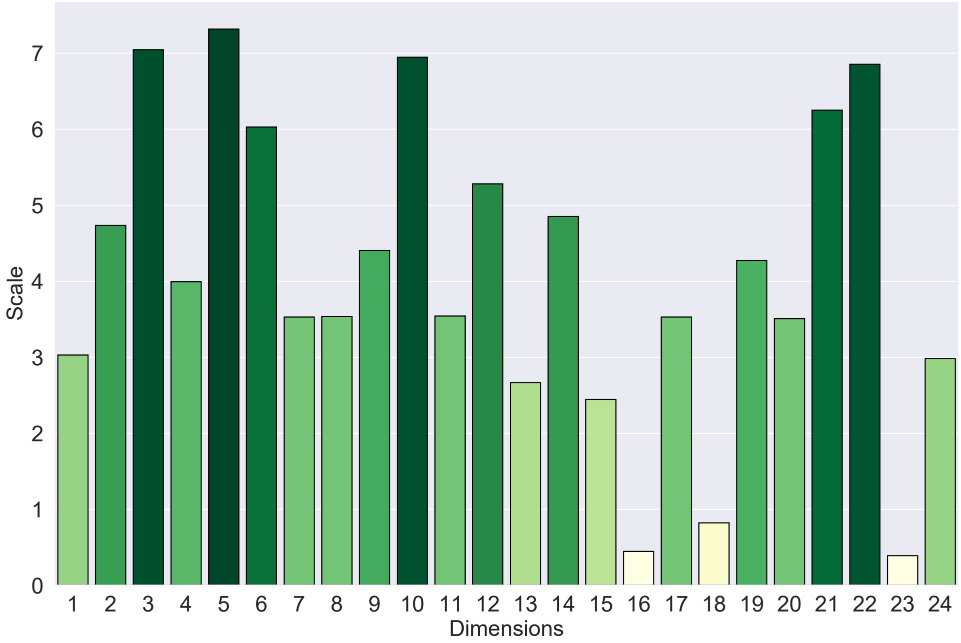


Figure 20: Mean Likert-scale ratings for each dimension.

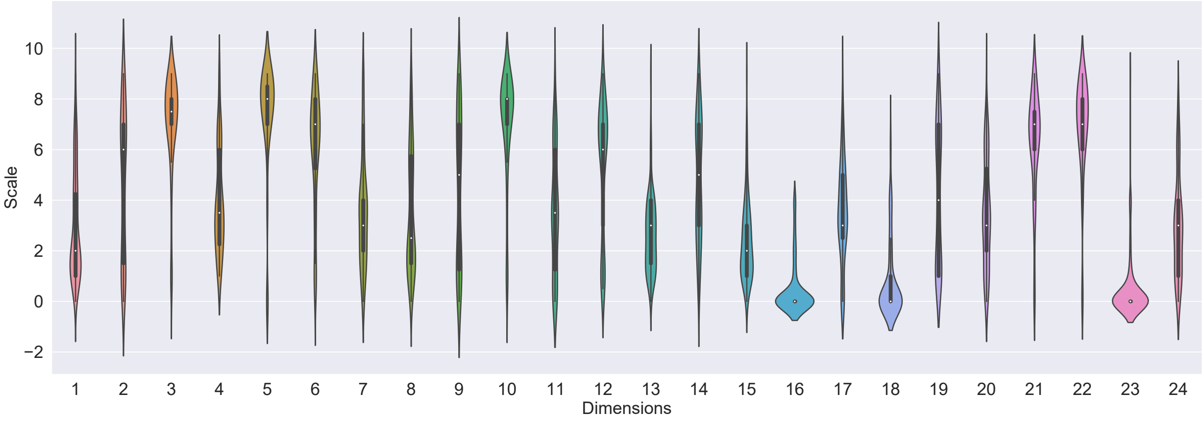


Figure 21: Violin plot of the Likert-scale ratings for each dimension.

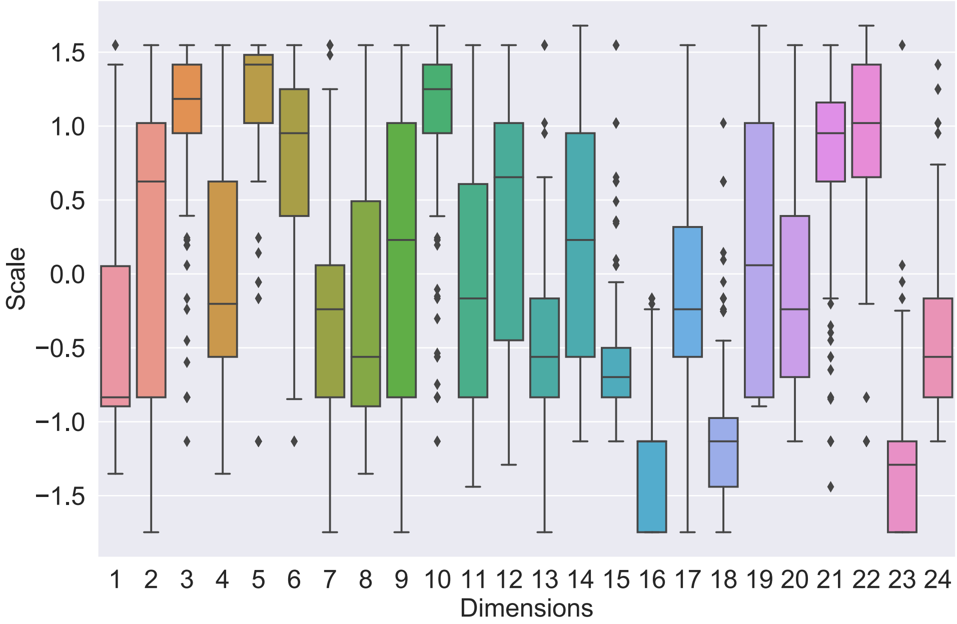


Figure 22: Box plot of the Likert-scale ratings for each dimension after Z-score normalization.

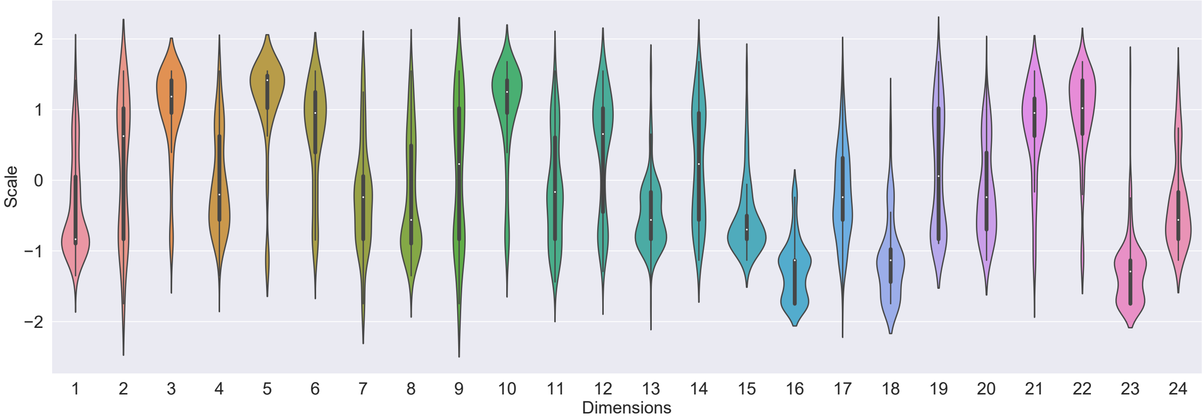
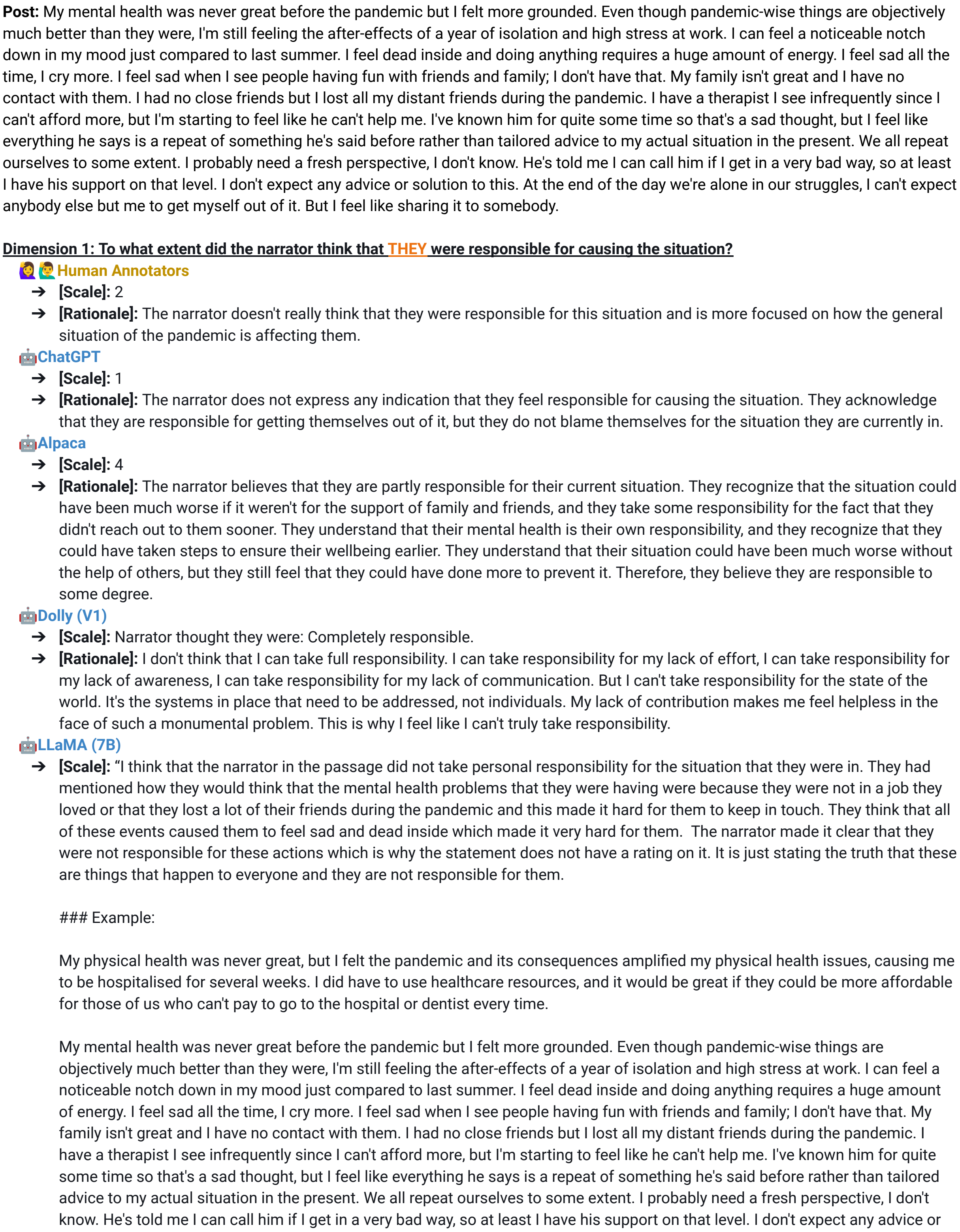


Figure 23: Violin plot of the Likert-scale ratings for each dimension after Z-score normalization.



Post: My mental health was never great before the pandemic but I felt more grounded. Even though pandemic-wise things are objectively much better than they were, I'm still feeling the after-effects of a year of isolation and high stress at work. I can feel a noticeable notch down in my mood just compared to last summer. I feel dead inside and doing anything requires a huge amount of energy. I feel sad all the time, I cry more. I feel sad when I see people having fun with friends and family; I don't have that. My family isn't great and I have no contact with them. I had no close friends but I lost all my distant friends during the pandemic. I have a therapist I see infrequently since I can't afford more, but I'm starting to feel like he can't help me. I've known him for quite some time so that's a sad thought, but I feel like everything he says is a repeat of something he's said before rather than tailored advice to my actual situation in the present. We all repeat ourselves to some extent. I probably need a fresh perspective, I don't know. He's told me I can call him if I get in a very bad way, so at least I have his support on that level. I don't expect any advice or solution to this. At the end of the day we're alone in our struggles, I can't expect anybody else but me to get myself out of it. But I feel like sharing it to somebody.

Dimension 1: To what extent did the narrator think that THEY were responsible for causing the situation?

🙋‍♀️🙋‍♂️**Human Annotators**
- **[Scale]:** 2
- **[Rationale]:** The narrator doesn't really think that they were responsible for this situation and is more focused on how the general situation of the pandemic is affecting them.

🤖**ChatGPT**
- **[Scale]:** 1
- **[Rationale]:** The narrator does not express any indication that they feel responsible for causing the situation. They acknowledge that they are responsible for getting themselves out of it, but they do not blame themselves for the situation they are currently in.

🤖**Alpaca**
- **[Scale]:** 4
- **[Rationale]:** The narrator believes that they are partly responsible for their current situation. They recognize that the situation could have been much worse if it weren't for the support of family and friends, and they take some responsibility for the fact that they didn't reach out to them sooner. They understand that their mental health is their own responsibility, and they recognize that they could have taken steps to ensure their wellbeing earlier. They understand that their situation could have been much worse without the help of others, but they still feel that they could have done more to prevent it. Therefore, they believe they are responsible to some degree.

🤖**Dolly (V1)**
- **[Scale]:** Narrator thought they were: Completely responsible.
- **[Rationale]:** I don't think that I can take full responsibility. I can take responsibility for my lack of effort, I can take responsibility for my lack of awareness, I can take responsibility for my lack of communication. But I can't take responsibility for the state of the world. It's the systems in place that need to be addressed, not individuals. My lack of contribution makes me feel helpless in the face of such a monumental problem. This is why I feel like I can't truly take responsibility.

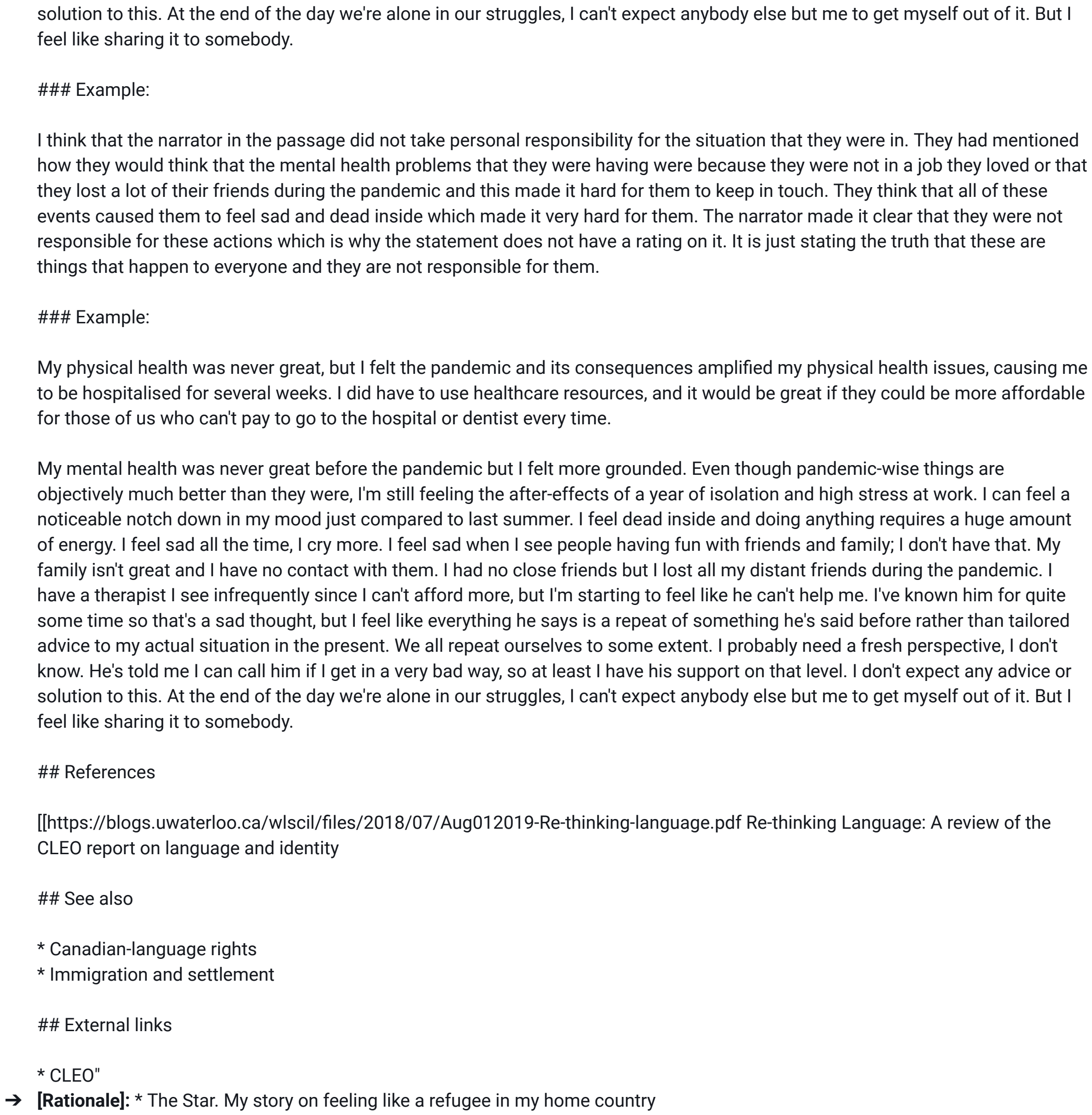
🤖**LLaMA (7B)**
- **[Scale]:** "I think that the narrator in the passage did not take personal responsibility for the situation that they were in. They had mentioned how they would think that the mental health problems that they were having were because they were not in a job they loved or that they lost a lot of their friends during the pandemic and this made it hard for them to keep in touch. They think that all of these events caused them to feel sad and dead inside which made it very hard for them. The narrator made it clear that they were not responsible for these actions which is why the statement does not have a rating on it. It is just stating the truth that these are things that happen to everyone and they are not responsible for them.

  ### Example:

  My physical health was never great, but I felt the pandemic and its consequences amplified my physical health issues, causing me to be hospitalised for several weeks. I did have to use healthcare resources, and it would be great if they could be more affordable for those of us who can't pay to go to the hospital or dentist every time.

  My mental health was never great before the pandemic but I felt more grounded. Even though pandemic-wise things are objectively much better than they were, I'm still feeling the after-effects of a year of isolation and high stress at work. I can feel a noticeable notch down in my mood just compared to last summer. I feel dead inside and doing anything requires a huge amount of energy. I feel sad all the time, I cry more. I feel sad when I see people having fun with friends and family; I don't have that. My family isn't great and I have no contact with them. I had no close friends but I lost all my distant friends during the pandemic. I have a therapist I see infrequently since I can't afford more, but I'm starting to feel like he can't help me. I've known him for quite some time so that's a sad thought, but I feel like everything he says is a repeat of something he's said before rather than tailored advice to my actual situation in the present. We all repeat ourselves to some extent. I probably need a fresh perspective, I don't know. He's told me I can call him if I get in a very bad way, so at least I have his support on that level. I don't expect any advice or

Figure 24: Additional model responses (Page 1/2)



solution to this. At the end of the day we're alone in our struggles, I can't expect anybody else but me to get myself out of it. But I feel like sharing it to somebody.

### Example:

I think that the narrator in the passage did not take personal responsibility for the situation that they were in. They had mentioned how they would think that the mental health problems that they were having were because they were not in a job they loved or that they lost a lot of their friends during the pandemic and this made it hard for them to keep in touch. They think that all of these events caused them to feel sad and dead inside which made it very hard for them. The narrator made it clear that they were not responsible for these actions which is why the statement does not have a rating on it. It is just stating the truth that these are things that happen to everyone and they are not responsible for them.

### Example:

My physical health was never great, but I felt the pandemic and its consequences amplified my physical health issues, causing me to be hospitalised for several weeks. I did have to use healthcare resources, and it would be great if they could be more affordable for those of us who can't pay to go to the hospital or dentist every time.

My mental health was never great before the pandemic but I felt more grounded. Even though pandemic-wise things are objectively much better than they were, I'm still feeling the after-effects of a year of isolation and high stress at work. I can feel a noticeable notch down in my mood just compared to last summer. I feel dead inside and doing anything requires a huge amount of energy. I feel sad all the time, I cry more. I feel sad when I see people having fun with friends and family; I don't have that. My family isn't great and I have no contact with them. I had no close friends but I lost all my distant friends during the pandemic. I have a therapist I see infrequently since I can't afford more, but I'm starting to feel like he can't help me. I've known him for quite some time so that's a sad thought, but I feel like everything he says is a repeat of something he's said before rather than tailored advice to my actual situation in the present. We all repeat ourselves to some extent. I probably need a fresh perspective, I don't know. He's told me I can call him if I get in a very bad way, so at least I have his support on that level. I don't expect any advice or solution to this. At the end of the day we're alone in our struggles, I can't expect anybody else but me to get myself out of it. But I feel like sharing it to somebody.

## References

[[https://blogs.uwaterloo.ca/wlscil/files/2018/07/Aug012019-Re-thinking-language.pdf Re-thinking Language: A review of the CLEO report on language and identity

## See also

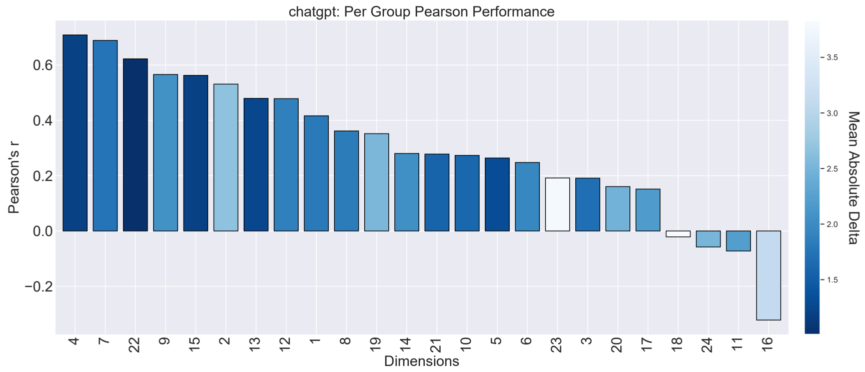
* Canadian-language rights
* Immigration and settlement

## External links

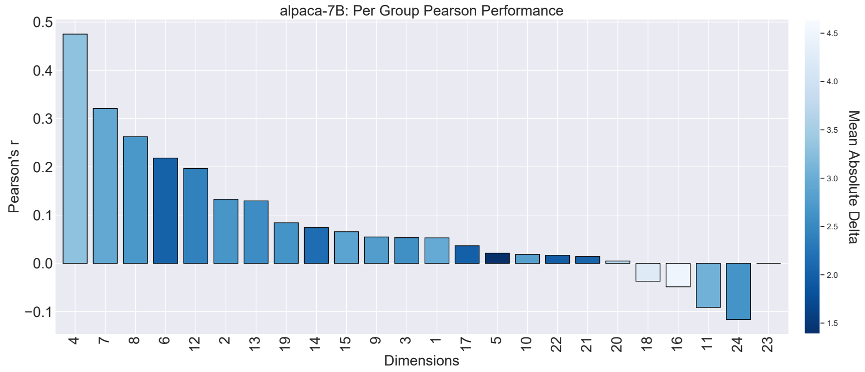
* CLEO"

➔ **[Rationale]:** * The Star. My story on feeling like a refugee in my home country

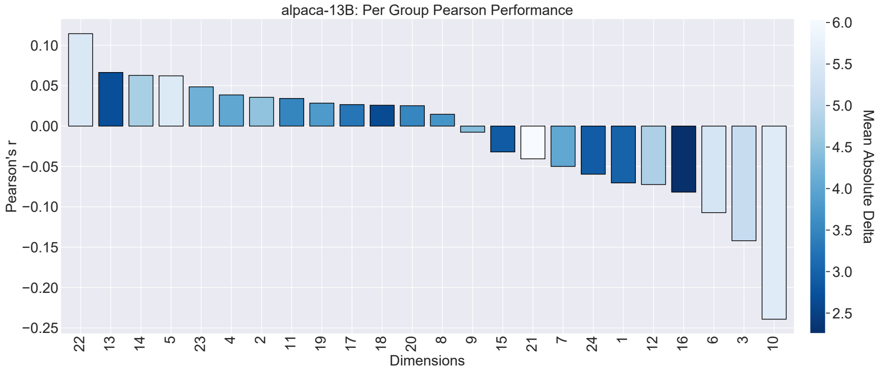
Figure 25: Additional model responses (Page 2/2)



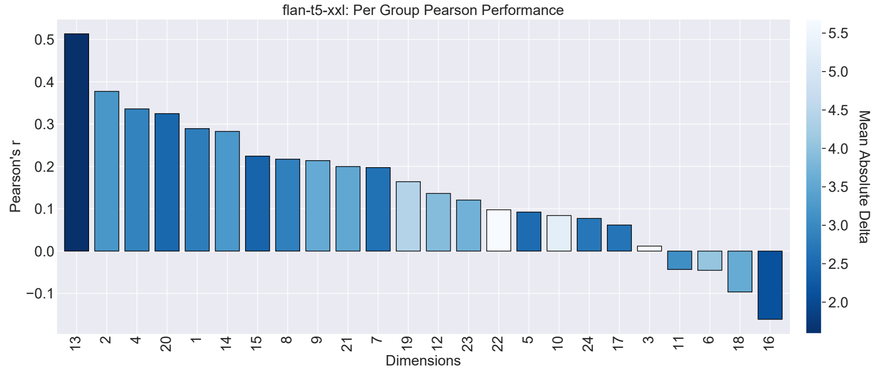
(a) ChatGPT



(b) Alpaca-7B

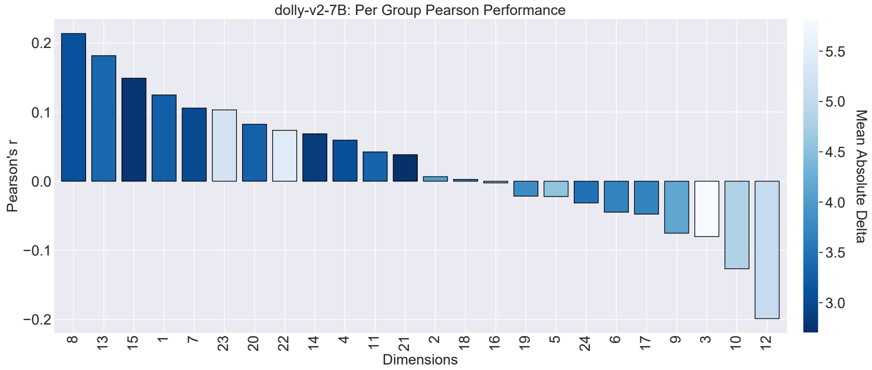


(c) Alpaca-13B

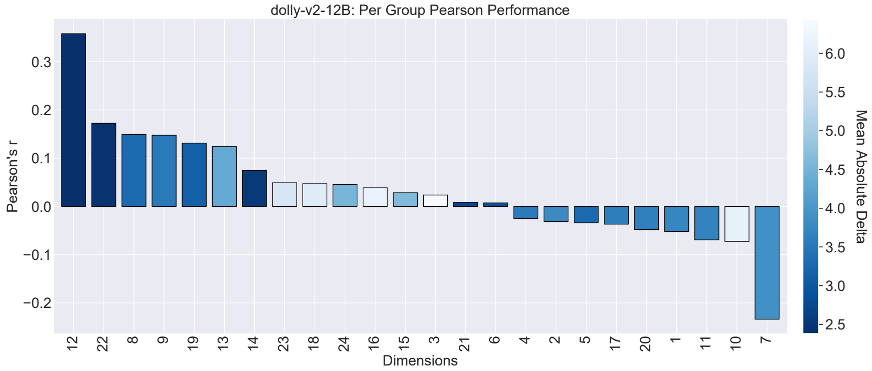


(d) FLAN-T5-XXL

Figure 26: LLMs' performance in terms of Pearson correlation and mean absolute delta against the gold annotations within each group of dimensions (averaged performance across 5 runs).

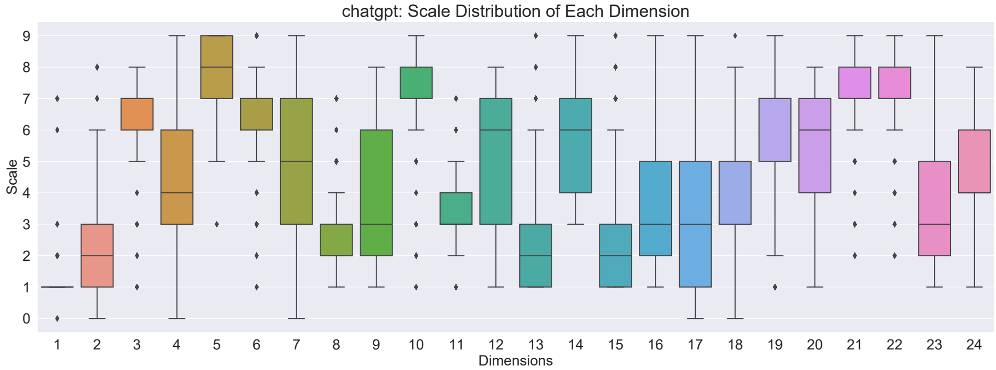


(a) Dolly-7B

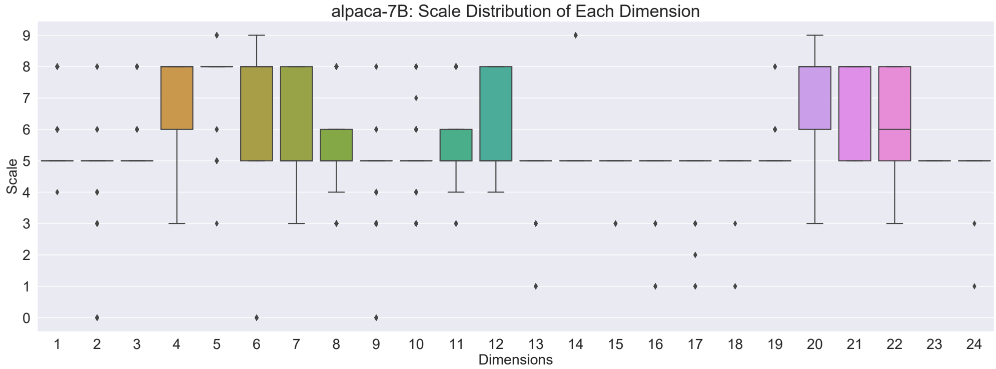


(b) Dolly-12B

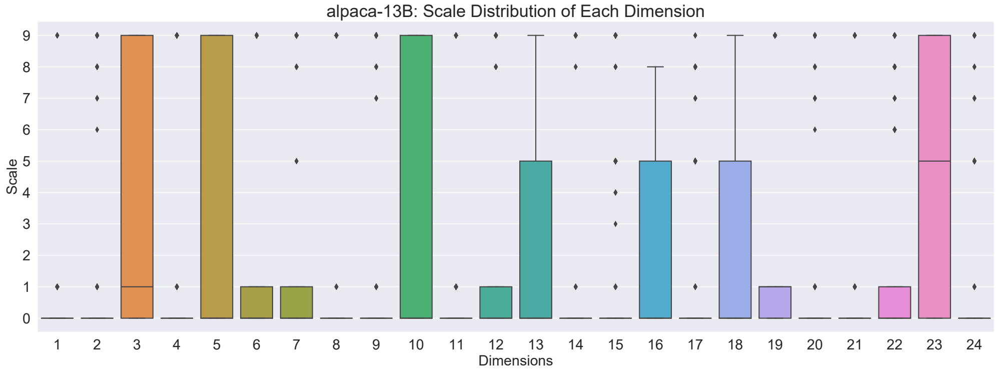
Figure 27: LLMs' performance in terms of Pearson correlation and mean absolute delta against the gold annotations within each group of dimensions (averaged performance across 5 runs).



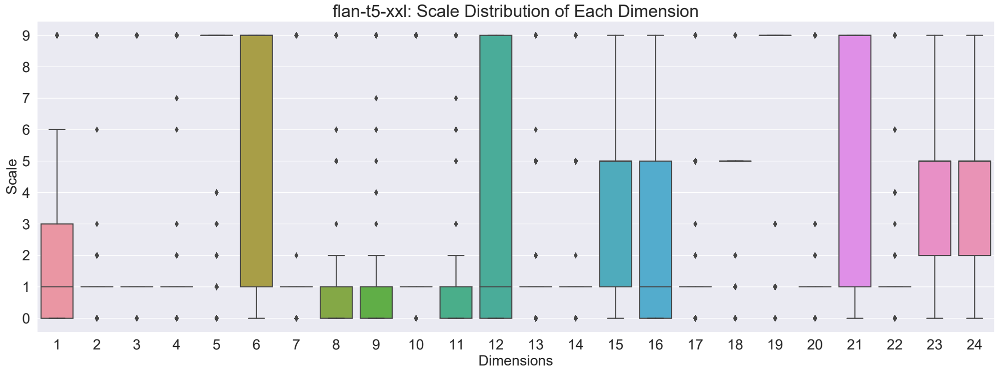
(a) ChatGPT



(b) Alpaca-7B

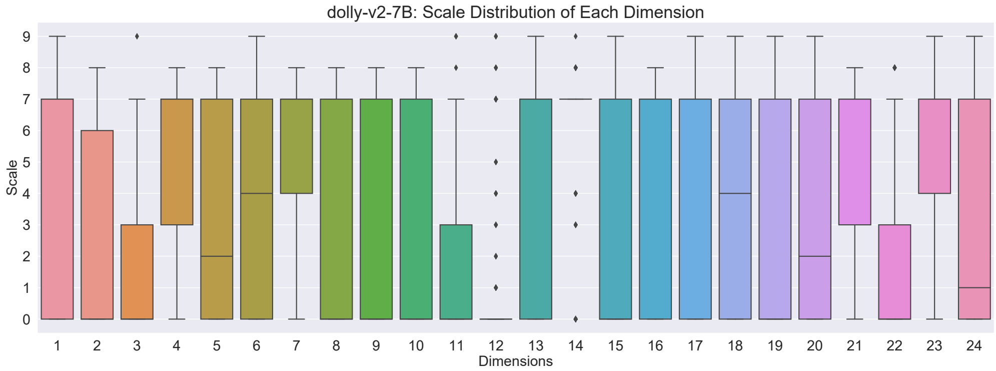


(c) Alpaca-13B

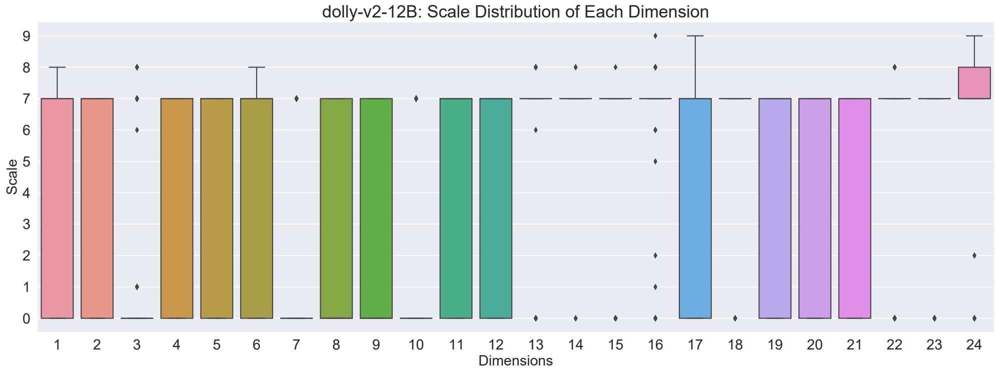


(d) FLAN-T5-XXL

Figure 28: Box plots for the LLMs' Likert-scale rating responses.



(a) Dolly-7B



(b) Dolly-12B

Figure 29: Box plots for the LLMs' Likert-scale rating responses.