# Modeling Subjectivity in Cognitive Appraisal with Language Models

**Yuxiang Zhou**[1,2*], **Hainiu Xu**[2*], **Desmond C. Ong**[3],
**Maria Liakata**[1,4], **Petr Slovak**[2], **Yulan He**[2,4]
[1]Queen Mary University of London, [2]King's College London
[3]The University of Texas at Austin, [4]The Alan Turing Institute
{yuxiang.zhou, m.liakata}@qmul.ac.uk
{hainiu.xu, petr.slovak, yulan.he}@kcl.ac.uk
desmond.ong@utexas.edu

## Abstract

As the utilization of language models in interdisciplinary, human-centered studies grow, expectations of their capabilities continue to evolve. Beyond excelling at conventional tasks, models are now expected to perform well on user-centric measurements involving confidence and human (dis)agreement- factors that reflect subjective preferences. While modeling subjectivity plays an essential role in cognitive science and has been extensively studied, its investigation at the intersection with NLP remains under-explored. In light of this gap, we explore how language models can quantify subjectivity in *cognitive appraisal* by conducting comprehensive experiments and analyses with both fine-tuned models and prompt-based large language models (LLMs). Our quantitative and qualitative results demonstrate that personality traits and demographic information are critical for measuring subjectivity, yet existing post-hoc calibration methods often fail to achieve satisfactory performance. Furthermore, our in-depth analysis provides valuable insights to guide future research at the intersection of NLP and cognitive science[1].

## 1 Introduction

Large language models (LLMs) are increasingly deployed in high-stakes scenarios such as healthcare (Sharma et al., 2023b) and law (Fan et al., 2024) where the integration of human oversight is essential to mitigate potential risks. However, designing such systems poses significant challenges, as real-world human decision-making is prone to occasional errors and subjectivity (Xiong et al., 2024; Zhou et al., 2023; Cheng and Vlachos, 2024). Moreover, inherent disagreements among human annotators, stemming from the subjective nature

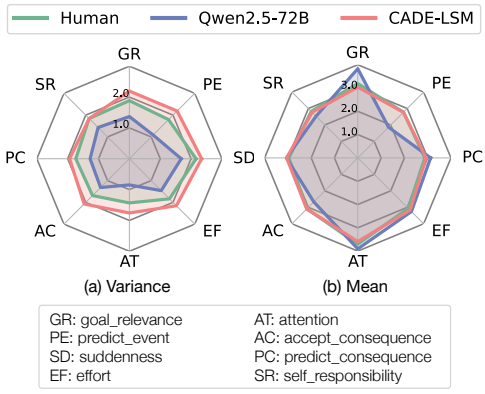


Figure 1: Illustration of mean and variance of appraisal distribution between human and different models on 8 appraisal dimensions (Table A1). While models exhibit adequate capability in modeling the mean (b), they are lacking in modeling the variance (a).

of their judgments, further complicate the process (Collins et al., 2023; Wang et al., 2024b). Most existing studies either assume the presence of a single human oracle, or aggregate multiple ratings using majority voting or averaging. However, such methods fail to capture the nuances of human subjectivity. Following Chen (2008), we define subjectivity as: *the property that creates variances in experiences, interpretations, or behaviors shaped by internal states and personal perspective- encompasses multiple dimensions such as uncertainty, vagueness, and imprecision.* A focus on subjectivity aims to understand and model the inherent variability reflected in individuals' internal states, which is fundamental to human reasoning about the world (O'Hagan et al., 2006; Lake et al., 2017; Chater et al., 2020) and is indispensable for designing responsible AI systems (Zhou et al., 2023; Cheng and Vlachos, 2024).

One important example of subjectivity is in how people experience and regulate emotions. Divergent emotional reactions may emerge from the same situation. In psychology, such variation is attributed to subjectivity in *cognitive appraisal*, the

[*]Equal contribution
[1]We make our code and resources available at https://github.com/seacowx/CogApp-LLM-Subjectivity.



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13811–13833
November 4-9, 2025 

process by which individuals evaluate and interpret events they experienced in relation to their beliefs, goals, and prior experiences. (Gross, 1998; Goldin et al., 2008; Giuliani and Gross, 2009; Skerry and Saxe, 2015; McRae, 2016; Yeo and Ong, 2024). For example, when experiencing a romantic breakup, some individuals may blame themselves and appraise the situation to be "self-responsible" (causing emotions like *guilt* or *regret*) while others may appraise it to be "others-responsible" (causing emotions like *anger*). Recent work has examined the ability of LLMs to identify subjective appraisals in how individuals evaluate their situations (Hofmann et al., 2020; Zhan et al., 2023; Yeo and Jaidka, 2023), which in turn shapes their emotional experiences. Despite the progress achieved by these studies, existing methods often fail to capture the nuanced variability among individuals. As illustrated in Figure 1, while models perform well in capturing average tendencies (mean), they struggle to model the inter-individual variance that is central to subjective interpretation. As a result, the models may end up representing only a narrow subset of individuals and lack the ability to properly model cognitive appraisal across varying population.

Modeling subjectivity enables user-centric evaluations that reflect individual preferences, which is crucial for the development of personalized and socially responsible systems. For instance, LLMs can be guided to generate tailored re-appraisals for emotion, grounded in estimated appraisal variance (Sharma et al., 2023b; Zhan et al., 2024). While this concept has been extensively studied in cognitive science (Schütz, 1942; Demszky et al., 2023; Sharma et al., 2023a), its investigation at the intersection with NLP remains underexplored (Abercrombie et al., 2024). To bridge this gap, we investigate how language models can be used to model subjectivity in cognitive appraisal. Motivated by the increasing demand for human-centred evaluation and the growing interdisciplinary applications of LLMs, we frame our study around three key research questions: 1) To what extent can language models quantify subjectivity in cognitive appraisal? 2) Can their ability to measure subjectivity be improved, and if so, how? 3) What insight can be gained from modeling subjectivity for practical applications? To address these questions, we conducted a series of experiments and analyses across various scenarios using both fine-tuned models and prompt-based LLMs. In summary, our contributions are as follows:

- We conducted a pilot study to investigate how language models can be utilized to model subjectivity in cognitive appraisal.
- We explored methods to improve subjectivity quantification from two perspectives: knowledge injection and post-hoc calibration. Our findings suggest that personal profiles[2] including personality traits and demographic information play a critical role in achieving better results, whereas current post-hoc calibration approaches often fail to produce satisfactory results.
- Our in-depth qualitative analysis provides valuable insights for future research at the intersection of NLP and cognitive science.

## 2 Related Work

**Subjectivity Modeling.** Existing NLP studies predominantly focus on modeling disagreement, aiming to reconcile conflicts among subjective viewpoints (Pang and Lee, 2004; Chen, 2008; Uma et al., 2021; Paun and Simpson, 2021; Plank, 2022; Aher et al., 2023). For example Wang et al. (2024b) accounted for agreements among humans to train a preference model for natural language generation. Leonardelli et al. (2021) focused on the conflict of human annotators and investigated the impact of different degrees of disagreement. However, subjectivity modeling seeks to leverage the inherent variability in individuals' internal states and personal perspectives, thereby capturing personal and context-dependent nuances required for practical applications (Shokri et al., 2024; Giorgi et al., 2024). The importance of subjectivity modeling stems not only from the growing deployment of language models in interdisciplinary research involving human-centred tasks and computational social science, but also from the need to ensure responsible and trustworthy human–AI interactions in high-stakes domains (Gordon et al., 2022; Giorgi et al., 2024). In light of this research gap, this paper explores foundational steps toward the development of language models capable of modeling subjectivity in cognitive appraisal.

**Appraisal in Psychology.** Understanding cognitive appraisal is essential for empowering individuals to regulate their emotions (Gross, 1998; Goldin et al., 2008; Giuliani and Gross, 2009; Skerry and Saxe, 2015; McRae, 2016). For instance, designing reappraisal interventions helps people change their

[2]**All personal profile information is anonymized and cannot be traced back to its provider.**

interpretation of situations that trigger undesirable emotions (Gross, 2015). Psychological research has been focused on investigating the underlying cognitive mechanisms of this process. Uusberg et al. (2019) proposed *reAppraisal* framework to explain these mechanisms in light of appraisal theory (Scherer, 2001). Building on this work, Uusberg et al. (2023) modeled the cognitive process involved in reappraisal aross various contexts by representing instances of reappraisal as profiles of shifts along abstract appraisal dimensions that characterize the significance of a situation for salient motives. In a recent meta-analysis, Yeo and Ong (2024) identified a comprehensive list of 47 cognitive appraisal dimensions studied in across various theories. In this paper, we followed Hofmann et al. (2020) to adapt 21 appraisal dimensions that have been widely examined in diverse scenarios in both psychological and NLP communities.

**Appraisal in NLP.** Early efforts in appraisal analysis focused on classifying the appraisal dimensions conveyed in a given situation (Hofmann et al., 2020), for example, identifying the most likely appraisal from pre-defined dimensions, such as *attention, coping, and responsibility*. Another line of research focused on estimating the strength of people's appraisals in response to a situation they are experiencing. For instance, Zhan et al. (2023) and Yeo and Jaidka (2023) have introduced datasets to evaluate how LLMs assess cognitive appraisals. Other studies have explored how LLMs can assist people in reframing negative appraisals. Sharma et al. (2023b) showed that highly specific and actionable appraisal reframing is considered the most helpful. Zhan et al. (2024) suggested that LLMs can be guided to generate reappraisal response for emotional support with psychological principles. Our work differs from previous studies in three key ways. First, rather than focusing on a single appraisal dimension, we assume that multiple appraisal profile are underlying the appraisal process. Second, instead of predicting point estimates, we model the underlying appraisal distribution, which allows us to capture subjectivity and variance. Third, we evaluate our model in various real-world scenarios and find that people's appraisal pattern differ across situations.

## 3 Modeling Subjectivity in Cognitive Appraisal Evaluation

In psychology, Cognitive Appraisal Evaluation (CAE), the task of predicting human appraisal judgments on a given situation, is typically conducted by asking individuals to provide Likert-scale ratings[3] on a set of predefined appraisal dimensions related to an event or situation (Smith and Ellsworth, 1985). Within the NLP community, researchers commonly frame CAE as a classification task, where human-provided appraisal ratings serve as ground truth labels (Zhan et al., 2023).

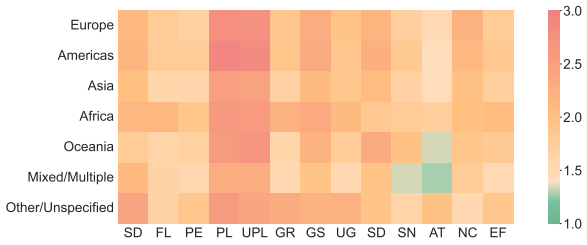


Figure 2: Appraisal variance across different geographic locations in the EnVent dataset (Hofmann et al., 2020). The $x$-axis represents the appraisal dimensions as defined by Hofmann et al. (2020), while the $y$-axis represents the origins of the participants. See § A1 for a detailed description of the appraisal dimensions.

However, numerous studies have shown that human appraisal ratings are not entirely consistent, even when assessing the same situation (Peacock and Wong, 1990; Troiano et al., 2023). As illustrated in Figure 2, individuals from different geographical regions exhibit varying appraisal ratings for the same scenarios. From a micro perspective, variability in certain appraisal dimensions differs across geographical groups. For example, the variance of *attention* (AT) among individuals from *Oceania* is significantly lower than that of individuals from *Africa*. From a macro perspective, the overall appraisal patterns also differ across geographical locations. For example, individuals from *Mixed/Multiple* locations tend to have higher consensus across different appraisal dimensions compared to those from the *Americas*.

To address the challenge of capturing the subjectivity inherent in human appraisal judgments, we adopt the Boltzmann policy for rating cognitive appraisals and propose a novel *distribution-estimation* task for CAE. Formally, given a situation, $s$, and an appraisal dimension, $\delta_i \in \Delta$, the goal is to model the underlying distribution of the rating, $R_{\delta_i}$, $\mathbb{P}(R_{\delta_i} = r_{\delta_i}|s)$. Assuming that the rating at $\delta_i$ follows some latent distribution $d \in \mathcal{D}$, parameterized by $\theta \in \Theta$ (i.e. $R_{\delta_i} \sim d(\theta)$), the

[3]The human rater could be the event experiencer themselves or an observer of the event.

distribution estimation task can be formulated as:

$$\hat{d}(\hat{\theta}) = \underset{d\in\mathcal{D},\theta\in\Theta}{\operatorname{argmin}} \operatorname{dist}\left(\overline{d(\theta)}_{\delta_i}, d(\theta)\right), \quad (1)$$

where $\operatorname{dist}(\cdot,\cdot)$ is some distance metric between two probability distributions, and $\overline{d(\theta)}_{\delta_i}$ is the sample distribution obtained from repeated human measurements. Minding that different $R_{\delta_i}$ may follow distinct distributions, we do not make any assumption on the distribution $d(\theta)$. Instead of parameter estimation, an unorthodox task for language models, we construct sample distributions using language models and directly evaluate their quality.

We utilize two sets of metrics to evaluate the estimated distribution, namely a set of point estimate metrics and a distribution metric.

**Point Estimate Metrics** To evaluate the quality of the modeled distribution, a convenient measure is to assess the estimated mean and variance. We use sample mean and sample variance obtained from repeated measurements as the ground truth and compute the Mean Absolute Error (MAE) for both the estimated mean, $\mu$-MAE, and the estimated variance, $\sigma^2$-MAE.

**Distribution Metric** While the point estimates provide assessments of the quality of the estimated mean or variance, they do not reflect the holistic quality of the estimated distribution. To comprehensively evaluate the discrepancy between the sample distribution and the modeled distribution, a common measure is the KL divergence. However, KL divergence suffers from asymmetry and vulnerability to regions with zero probability density. As such, we opt to use the Wasserstein distance, which is a proper metric that measures the distance between two probability distributions using optimal transport plan (Kantorovich, 1960). Formally, given two discrete probability distributions $p$ and $q$, the Wasserstein distance[4] is defined as

$$W_1(p,q) = \int_{\mathbb{R}} |F_p(x) - F_q(x)|dx \quad (2)$$

where $F_p(x)$ and $F_q(x)$ are the cumulative distribution function of $p$ and $q$ respectively. We report the average Wasserstein distance across all appraisal dimensions as the final evaluation score.

## 4 Preliminary Explorations

To investigate language models' capability in modeling the subjectivity in CAE, we examine two categories of methods: fine-tuning pre-trained autoencoding language models (PLMs) and zero-shot prompting of autoregressive LLMs.

### 4.1 Fine-tuning of Pre-trained Models

To tackle the distribution modeling task, we leverage the base model in two approaches, namely *label-smoothing* and *variational inference*.

**Label Smoothing** Originally introduced by Szegedy et al. (2016) as a regularization method, *Label Smoothing* has since been widely adopted in classification tasks. It reduces the kurtosis of the output logits by disseminating a portion of the probability density from ground truth label. In distribution modeling, label densities ought to be allocated so that they resemble the the sample distribution. However, such a smoothing approach is infeasible as obtaining a sample distribution would require a large number of repeated measurements. Recognizing that 64% of the appraisal ratings follow a unimodal distribution and 35% follow a bimodal distribution[5], we examine two label smoothing approaches. Firstly, we apply label smoothing using a discretized Gaussian distribution centered at the ground truth rating, which effectively reflect the unimodal distribution of appraisal ratings. Secondly, we examine label smoothing with a mixture of two discretized Gaussian distributions, which replicate the bimodal distribution[6]. We fine-tune PLMs using the smoothed labels as the target distribution. We refer to this approach as **CADE-LSM**.

**Variational Inference** The task of distribution-estimation has been widely studied in the machine learning community through variational inference (Jordan et al., 1999; Wainwright et al., 2008). From a Bayesian perspective, estimating the distribution in CAE is equivalent to inferring the posterior distribution of appraisal ratings given a situation. To this end, we adopt a Variational Autoencoder (VAE) (Kingma and Welling, 2013) to model the posterior distribution. Specifically, we use a PLM as the encoder to estimate the parameters of the latent distribution. Samples from the reparameterized latent distribution are then decoded using a two-layer MLP network to compute the evidence lower bound (ELBO). We additionally compute the Mean Squared Error (MSE) loss between the predicted

[4] We use the Wasserstein-1 distance.

[5] See § D for detailed analysis of modalities of appraisal rating distributions

[6] Due to the subpar performance of the bimodal label smoothing approach, we defer the details to §F.

and the ground truth appraisal ratings. The model is trained using the sum of the ELBO and the MSE loss. We refer to this approach as **CADE-VAE**. Implementation details are in §H.

### 4.2 Zero-shot Prompting

Advancements in scaling autoregressive LLMs and aligning them with human preference data have enabled LLMs to mimic human communication and reasoning. As such, we explore whether LLMs can be used to model the subjectivity in CAE via zero-shot prompting. We investigate this under two settings. In the first setting, LLMs are tasked to provide an appraisal rating given a description of the situation and a description of a specific appraisal dimension[7]. For each situation, we prompt LLMs 30 times to obtain a sampled distribution of appraisal ratings. In the second setting, we additionally provide the LLMs with a personalized description, which include the Big-Five personality traits as well as demographic information. In this case, we prompt LLMs with the same situation paired with different personal profile to obtain a sampled distribution of appraisal ratings. Details of the prompts are provided in §E.

**Finding the Optimal Temperature** When sampling appraisal ratings from LLMs, a key factor is the `"temperature"` parameter, which controls the randomness of the token generating process. To find an optimal temperature, we conduct a grid search over the temperature range $[0, 1.5]$. We select the temperature that yields the sampled distribution that is most similar to the distribution obtained from human ratings (§ A5).

**Post-hoc Calibration** There has been bountiful studies that look into post-hoc methods for calibrating LLMs' confidence in their predictions (Lin et al., 2022; Tian et al., 2023). While these methods emphasize the calibration of confidence (probability density) on the correct answer instead of the distribution over the label space, we examine whether these methods can be adopted to calibrate the distribution over all possible appraisal ratings. Specifically, we follow Xiong et al. (2024) and test two calibration methods: *Average Confidence* and *Pair Ranking* (§5.1).

[7]We follow (Hofmann et al., 2020) to formulate the question for each appraisal dimension.

## 5 Experiments

To address our research questions, we systematically evaluate two categories of methods across three datasets: EnVent dataset (Hofmann et al., 2020) consists of daily event descriptions produced by native English speakers with 21 annotated appraisal dimensions. FGE dataset (Skerry and Saxe, 2015) includes descriptions of emotion-eliciting events with annotations along 38 appraisal dimensions. CovidET dataset (Zhan et al., 2023) consists of situations described in Reddit posts related to COVID-19 and annotated along 24 appraisal dimensions. Detailed statistics of the datasets can be found in Table 1, and the full list of appraisal dimensions can be found in §A. Further details on annotation validity can be found in §C.

| **Dataset** | Size | Avg Len | # App Dim | # Ann |
|---|---|---|---|---|
| EnVent | 1,200 | 111.2 | 21 | 5 |
| FGE | 200 | 291.1 | 14 | 8 |
| CovidET | 40 | 727.3 | 16 | 2 |

Table 1: Statistics of the evaluation datasets: *Size*: dataset size / number of events; *Avg Len*: average length of situation description in words; *App Dim*: the number of appraisal dimensions; *Ann*: the number of human annotators for each event.

**Setup Details** For fine-tuned models, we use DeBERTa-V3-Large as the backbone (He et al., 2023). Each model is trained for 30 epochs with a linear warmup for the first 10% of the training steps. We employ AdamW (Loshchilov and Hutter, 2019) as the optimizer. We set the maximum learning rate at 5e-5 with a batch size of 32 and select the optimal model weights based on MSE loss on the development set. For in-domain experiments, we train on 4,680 examples from the EnVent dataset and select the best checkpoint using 540 validation examples. For out-of-domain experiments, we fine-tune on EnVent training set and evaluate on the FGE and CovidET datasets. Since the appraisal dimensions used in out-of-domain datasets slightly differ from those in EnVent, we manually aligned their dimensions and ratings for fair comparision. Further details are provided in §A.

**Baselines** The baseline models selected for comparison can be broadly categorized into two groups: fine-tuned models: **CADE-LSM** estimates subjectivity using label-smoothing (Rolf et al., 2022; Wang et al., 2024b); **CADE-VAE**, a latent variable model infers appraisal distribution using VAE (Kingma and Welling, 2013).

| Models | EnVent | | | FGE | | | CovidET | | |
|---|---|---|---|---|---|---|---|---|---|
| | Wasserstein ↓ | $\mu$-MAE ↓ | $\sigma^2$-MAE ↓ | Wasserstein ↓ | $\mu$-MAE ↓ | $\sigma^2$-MAE ↓ | Wasserstein ↓ | $\mu$-MAE ↓ | $\sigma^2$-MAE ↓ |
| Random | 1.196 | 1.096 | 1.060 | 1.191 | 1.088 | 1.042 | 1.438 | 1.367 | 0.833 |
| Majority | 1.392 | 1.275 | 0.883 | 1.313 | 1.222 | 0.775 | 0.950 | 0.918 | 0.332 |
| CADE-VAE | 1.279 | 0.984 | 0.882 | 1.209 | 1.105 | 0.713 | 1.200 | 1.106 | 0.331 |
| CADE-LSM | **0.773** | **0.665** | 0.837 | **0.926** | **0.835** | 0.795 | 1.112 | 1.023 | 0.642 |
| Llama3.1-8B | 1.094 | 0.904 | 0.826 | 1.409 | 1.385 | 0.780 | 1.109 | 1.065 | 0.361 |
| Qwen2.5-7B | 1.078 | 0.919 | 0.817 | 1.131 | 1.020 | 0.704 | **0.905** | **0.864** | **0.322** |
| Llama3.3-70B | 1.012 | 0.926 | 0.820 | 1.248 | 1.203 | 0.740 | 0.970 | 0.924 | 0.325 |
| Qwen2.5-72B | 0.945 | 0.852 | **0.736** | 1.048 | 0.960 | **0.632** | 0.905 | 0.867 | 0.325 |

Table 2: Main Results for the estimation of appraisal rating distribution. For all metrics, lower is better.

Prompt-based models include: **Llama3.1-8B** and **Llama3.3-70B** (AI@Meta, 2024); **Qwen2.5-7B** and **Qwen2.5-72B** (QwenTeam, 2024).

### 5.1 Results and Analysis

**How well do language models quantify subjectivity in cognitive appraisal ratings?**

***Fine-tuning PLMs brings significant improvement in mean estimation.*** As discussed in §4, we use the Wasserstein distance as the primary metric for its comprehensiveness. Point estimate metrics are utilized to provide insights into two aspects of the estimated distribution: mean and variance.

As shown in Table 2, fine-tuning on the EnVent dataset with label smoothing consistently outperforms all baseline models in terms of Wasserstein distance on both EnVent and FGE dataset. For instance, CADE-LSM surpasses the second-best baseline, Qwen2.5-7B, by 39.5% in Wasserstein distance. However, when examining the point estimate of the mean and variance, we see that the improvement in Wasserstein distance is largely attributed to the improvement in mean estimation. In terms of variance estimation, CADE-LSM underperforms all LLMs, suggesting that while fine-tuning PLMs can effectively capture the *general tendency* of human appraisal ratings (as reflected in the mean), it struggles to model the subjectivity in appraisal ratings (as reflected in the variance).

Moreover, fine-tuned models suffer from limited generalizability. For instance, trained with EnVent, in which the situations are described with concise sentences, CADE-LSM generalizes well to the FGE datasets, which contains situation descriptions of similar kind[8], but underperforms the Majority baseline on the CovidET dataset, which contains reddit posts that are longer in length and focus on COVID-19 related events.

[8]EnVent and FGE datasets contain daily event descriptions.

***LLMs are more effective in modeling variance in appraisal ratings.*** Although LLMs are less effective compared to fine-tuned PLMs in terms of the Wasserstein distance, they are more adept at modeling the variance in appraisal ratings. As shown in Table 2, the Qwen2.5 family consistently achieves the best point estimate for variance ($\sigma^2$-MAE). Similar to the variance-bias tradeoff, LLMs are good at modeling the variance while failing at reducing the bias. However, such results are insufficient in proving LLMs' effectiveness in modeling variance since the randomness of LLM generations can be easily adjusted by altering the temperature parameter (Table A5). To obtain a concrete conclusion, results in the following controlled experiments show that incorporating personal profile in the prompt significantly improves the estimated variance, especially for the Qwen2.5 family of models (Table 3). We also observe differences in behavior with respect to model scale. In general, large models achieved stronger overall performance. However, these improvements were not uniformly distributed, and no single model consistently outperformed others across all settings and datasets.

**Does personal profile help language models capture subjectivity?**

Studies in psychology have shown that personal profiles such as personality traits are critical to shaping human appraisal judgement (Mischel and Shoda, 1995; Childs et al., 2014). To understand the role of personal profile in modeling subjectivity with language models, we incorporate two types of information: demographic information and the Big-Five personality traits (Gosling et al., 2003).

***Effectiveness of personal profile varies across models.*** As shown in Table 3, while models incorporated with personal profile generally achieve better performance in terms of Wasserstein distance, the degree of performance increment is model-

dependent. The introduction of personal profiles generally leads to better performance for lLMs compared to fine-tuned PLMs, although the latter still achieve marginal improvements in certain scenarios. Specifically, we observe that adding personal profile is particularly effective for Qwen2.5-7B, which shows improvement across all metrics. In contrast, Llama3-8B only shows slight improvements in Wasserstein distance and $\sigma^2$-MAE, but decreased performance in $\mu$-MAE.

***Different types of personal profile contribute differently to model performance.*** To understand the contribution of each type of personal profile, we examine all models with either demographic (e.g., gender) information or personality traits (e.g., extraversion). As shown in Table 3, we found that adding personal profile information brings contrasting effects on fine-tuned models: while CADE-VAE does not benefit from the additional information, CADE-LSM is able to better capture the subjectivity in the appraisal ratings, which is evident by its improved *Wasserstein* and $\sigma^2$-MAE. For prompt-based models, results are mixed. Incorporating personality traits or demographic information improves mean modeling ($\mu$-MAE) for all LLMs except Llama3-8B, but this does not consistently yield better *Wasserstein* scores, as observed with Qwen2.5-72B.

***Significant improvement from personal profiles integration.*** To evaluate the impact of incorporating personal profiles on the subjectivity of various appraisal dimensions, we conducted a one-tailed two-sample T-test. Comprehensive p-value results are shown in Table A8. Our results show that 1/3 of the appraisal dimensions exhibit statistically significant improvements when either personality traits or demographical information are integrated, across both Llama3.1-8B and Qwen2.5-7B. For example, when both types of personal profiles are integrated, significant gains are observed for *predict_event, self_control, accept_consequence*, and *effort*. Detailed analyzes can be found in §I. These findings reassure that incorporating personal profiles play a vital role in modeling subjectivity.

**Do existing post-hoc calibration methods improve the modeling of subjectivity?**

***Post-hoc calibration degrades model performance.*** We conducted experiments using Avg-Conf and Pair-Rank methods (Xiong et al., 2024). In the setting of CAE, Avg-Conf samples multiple (rating confidence) pairs from LLMs by explicitly instructing LLMs to output a confidence score associated with each rating. Calibration is done by normalizing the confidence scores across all ratings. Pair-Rank assumes that LLMs are better at ranking a given set of ratings. By sampling multiple rankings of the ratings, pair-rank computes a categorical distribution over the rating space using stochastic gradient descent to optimize a conditional log-likelihood function based on the sampled rankings.

| Models | Wasserstein ↓ | $\mu$-MAE ↓ | $\sigma^2$-MAE ↓ |
|---|---|---|---|
| CADE-VAE | 1.279 | 0.984 | 0.882 |
| w. Demo | 1.281 | 0.983 | 0.882 |
| w. Traits | 1.281 | 0.983 | 0.882 |
| w. Demo & Traits | 1.279 | 0.984 | 0.882 |
| CADE-LSM | 0.773 | 0.665 | 0.837 |
| w. Demo | 0.768 | 0.682 | 0.843 |
| w. Traits | 0.763 | 0.676 | 0.832 |
| w. Demo & Traits | 0.765 | 0.678 | 0.818 |
| Llama3-8B | 1.094 | 0.904 | 0.826 |
| w. Demo | 1.090 | 1.014 | 0.823 |
| w. Traits | 1.148 | 1.078 | 0.848 |
| w. Demo & Traits | 1.086 | 1.008 | 0.817 |
| Qwen2.5-7B | 1.078 | 0.919 | 0.817 |
| w. Demo | 0.994 | 0.845 | 0.799 |
| w. Traits | 0.962 | 0.821 | 0.775 |
| w. Demo & Traits | 0.958 | 0.826 | 0.756 |
| Llama3-70B | 1.012 | 0.926 | 0.820 |
| w. Demo | 0.992 | 0.872 | 0.828 |
| w. Traits | 0.979 | 0.871 | 0.805 |
| w. Demo & Traits | 0.967 | 0.841 | 0.811 |
| Qwen2.5-72B | 0.945 | 0.852 | 0.736 |
| w. Demo | 0.961 | 0.838 | 0.738 |
| w. Traits | 0.946 | 0.828 | 0.724 |
| w. Demo & Traits | 0.948 | 0.836 | 0.714 |

Table 3: Personal profile study with various models on EnVent dataset. Improved, Unchanged, and Decreased results are highlighted.

| Models | Wasserstein ↓ | $\mu$-MAE ↓ | $\sigma^2$-MAE ↓ |
|---|---|---|---|
| Llama3-8B | 1.094 | 0.904 | 0.826 |
| w. Avg-Conf | 1.127 | 1.022 | 0.804 |
| w. Pair-Rank | 1.768 | 1.752 | 0.788 |
| Qwen2.5-7B | 1.078 | 0.919 | 0.817 |
| w. Avg-Conf | 1.045 | 0.873 | 0.828 |
| w. Pair-Rank | 1.167 | 1.081 | 0.797 |

Table 4: Post-hoc calibration on EnVent dataset. Lower values are better. Improved and Decreased results are highlighted in corresponding color.

Results shown in Table 4 demonstrate that the Avg-Conf and Pair-Rank calibration strategies lead to limited improvement in the case of Qwen2.5-7B and degraded performance in Llama3-8B. One possible explanation is that existing calibration meth-

| Models | Top Quantified Appraisal Dimensions | | |
|---|---|---|---|
| | Rank 1 | Rank 2 | Rank 3 |
| CADE-LSM | *pleasantness* | *unpleasantness* | *social_norms* |
| Lama3-8B | *social_norms* | *pleasantness* | *unpleasantness* |
| Qwen2.5-7B | *pleasantness* | *social_norms* | *unpleasantness* |
| Llama3.3-70B | *pleasantness* | *unpleasantness* | *social_norms* |
| Qwen2.5-72B | *social_norms* | *suddenness* | *effort* |
| | Bottom Quantified Appraisal Dimensions | | |
| CADE-LSM | *urgency* | *goal_support* | *predict_consequence* |
| Llama3.1-8B | *other_control* | *goal_support* | *familiarity* |
| Qwen2.5-7B | *familiarity* | *goal_support* | *other_control* |
| Llama3.3-70B | *other_control* | *attention* | *accept_consequence* |
| Qwen2.5-72B | *familiarity* | *urgency* | *self_responsibility* |

Table 5: Qualitative analysis for different models on various appraisal dimensions. Wasserstein distance is used as the measurement metric.

ods are designed for classification setup, where the objective is to align the logit of a predicted class with the probability of that the class being correct (Guo et al., 2017). This setup is fundamentally different from the distribution estimation task in CAE, where the probability density allocated to each rating must be modeled, which is much more intricate than the classification setup. Further research is needed to investigation the underlying reasons.

**Qualitative Study: Understanding Subjectivity**

Table 5 presents a qualitative analysis of subjectivity measurements for different models across various appraisal dimensions in the EnVent dataset. We observe slight differences in the top-three and bottom-three dimensions between fine-tuned and prompt-based models. For instance, CADE-LSM aligns most closely with human subjectivity on the dimensions of *pleasantness* and least on *urgency*, whereas Lama3.1-8B shows the strongest alignment on *social_norms* and the weakest on *other_control*. Nonetheless, we found that the dimensions of *pleasantness*, *social_norms*, and *unpleasantness* are captured most effectively by these models, while *familiarity*, *goal_support*, and *other_control* remain challenging for all models. Interestingly, we found that personal profile does not affect the qualitative results of subjectivity across all models. Comprehensive experiments are provided in § J. These findings highlight the need for future research to accurately quantify various dimensions of subjectivity.

**Effects of Demographics** To gain further insight on the effects of different demographic on the subjectivity of cognitive appraisal, we inspect the variance in different geographic locations in the EnVent dataset in Figure 3. The $x$-axis represents the appraisal dimensions defined in Hofmann et al. (2020). The $y$-axis represents the origins of the participants. As shown in Figure 3, CADE-LSM exhibits subjectivity patterns that align more closely with human judgments across all appraisal dimensions, suggesting substantial disagreement among individuals of different geographic locations in each dimension. Specifically, individuals show considerable variability in dimensions such as *pleasantness* (PL), *unpleasantness* (UPL), *self_responsibility* (SR), and *other_responsibility* (OR). In contrast, the Qwen2.5-72B model provides more consistent subjectivity quantification for individuals from different geographic locations in dimensions like *predict_event* (PE) and *predict_consequence* (PC). We deduce that, compared to prompt-based LLMs, fine-tuning allows the model to learn contextualized personal profile, thereby enabling them to more accurately capture subjectivity.

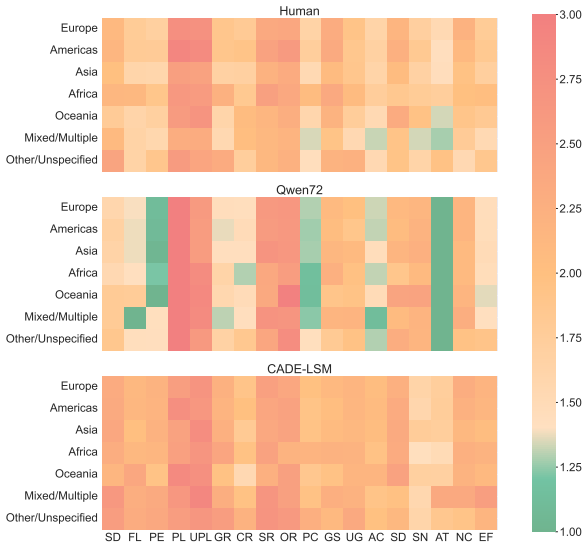


Figure 3: Comparison of appraisal variance in different geographic locations in the EnVent dataset. See Table A for definition of the abbreviated appraisal dimensions.

Upon further investigation, our study indicates that individuals with various geographic locations exhibit relatively consistent subjective cognition regarding *social_norms* (SN) and *familiarity* (FL), yet responsibility-related appraisal dimensions (e.g., CR, SR, OR) provoke more pronounced variability.

**Effects of Personality Traits** We conduct a similar analysis focusing on personality traits. We again find that fine-tuned models exhibit appraisal patterns closely align with those of human, indicating the critical role of personal profile in modeling subjectivity with fine-tuned models. Figure A3 provides further details on the comparison.

In summary, we found that certain appraisal di-

mensions such as *pleasantness* and *unpleasantness* show inherent variance regardless of individuals' geographic locations or personality traits. In contrast, dimensions such as *attention* and *accept_consequence* exhibit relatively small variations. We hypothesize that sentiment-related appraisal dimensions tend to elicit more extreme ratings, reflecting higher subjectivity than other dimensions. This effect may be driven by the annotation strategy used in EnVent dataset (Hofmann et al., 2020), which naturally incorporates bi-polar sentiment and thus amplifies variance. More discussion can be found in §J. Further research is needed to investigate the underlying factors driving these findings and to develop more advanced techniques for improving models' ability to model subjectivity in cognitive appraisal.

### Practical Implications

While mean-based measurements capture how representative a model is for that population, variance reflects how subjective a particular dimension may be within a population. In our experiments, we found that only week correlations between these two types of metics, and the relationship was model-dependent (§K). This suggests that beyond building models that representative of a population, it is equally important to develop models capable of capturing the nuanced subjectivity within each population. Building on our exploration of first two research question, we now provide practical recommendations for measuring subjectivity that are both psychologically meaningful and computationally feasible.

**Computational perspective** First, the metric selection depends on either task or dataset. For example, datasets share similar domain and event contexts tend to yield similar ranges of $\sigma$-MAE scores, vice versa. However we believe that the Wasserstein distance serves as a holistic measure for evaluating whether LLMs can approximate human judgement distributions in subjectivity tasks, as it captures both mean and variance. In practice, a model that poorly estimates the mean will also fail to represent the distribution meaningfully. Second, model choice depends on data availability. When sufficient annotated data is avaiable, we recommend fine-tuned models such as CADE-LSM, which perform better at holistic measurement of subjectivity. In low-resource setting, where annotations are sparse or unavailable, prompt-based LLMs can serve as a reasonable proxy with better generalizability, specifically for dimensions that benefit from personal profile integraiton, such as *predict_event, self_control*, and *accept_consequence*.

**Psychological Perspective** Certain appraisal dimensions are well-modeled by either PLMs or LLMs. Dimensions such as *pleasantness, unpleasantness*, and *self_responsibility* show high variance among annotators, yet LLMs are able to capture their subjectivity reasonably well. By contrast, *goal_support* also exhibits high variance, but models struggle to replicate it. In the same time, *social_norms* displays low variance in human ratings, and this stability is likewise well modeled by LLMs. However, *predict_consequence* shows low human variance but is poorly captured by models, suggesting that low variance alone does not guarantee the replicability.

## 6 Conclusion

In this paper, we take an important first step toward modeling subjectivity in cognitive appraisal with language models. We conduct detailed experiments and analysis across various scenarios with both fine-tuned PLMs and prompt-based LLMs. Our findings reveal notable inconsistencies in modeling subjectivity, with no single model reliably identifying it across all appraisal dimensions under varying conditions. The thorough quantitative and qualitative examination indicates that personality traits and demographical information play a vital role in measuring subjectivity, whereas existing post-hoc calibration approaches fail to achieve improved results. Furthermore, our qualitative analysis provides valuable insights for future research and development in the intersection of NLP, Cognitive Modeling, and Psycholinguistics.

## 7 Limitation

While this paper lays the groundwork for further exploration into subjectivity modeling, it is necessary to acknowledge its limitations. First, we treated appraisal dimensions independently, without consideration of their potential interrelationships. However, certain dimensions are closely correlated. For example, *self_control*, *other_control*, and *chance_control* are all fall in the coping objective addressing control attributions (Sander et al., 2005; Scherer and Fontaine, 2013; Troiano et al., 2023). By incorporating prior knowledge of cor-

relations among these dimensions, the appraisal distribution could be more accurately modeled for subjectivity. Secondly, the datasets used in our experiments involve a limited number of annotators, particularly the CovidET dataset, which included only two annotators per sample. We observed that in datasets with a moderate number of annotators (e.g., FGE and EnVent, where $n \geq 5$), model predictions exhibited similar trends, whereas the CovidET dataset showed different model behavior. To enhance statistical robustness, future work should incorporate a larger pool of annotators (e.g., $n \geq 30$). Thirdly, we employed relatively simple prompts for LLMs, whereas prior research suggests that more advanced prompting methods could enhance model performance (Wei et al., 2022; Zhou et al., 2024). Moreover, our experiments focused primarily on English, meaning the findings may not necessarily generalize to multilingual contexts (Banea et al., 2008, 2010). Investigating subjectivity modeling across different languages remains an avenue for future research. Finally, LLMs exhibit biases that can lead to unfaithful appraisal ratings, especially when persona profiles are involved (Wang et al., 2024a; Dong et al., 2024). Further work is needed to better align LLMs with provided profiles to ensure faithful and unbiased appraisal ratings.

## 8 Ethical Considerations

The primary goal of this study is to draw attention of quantifying inherent subjectivity in cognitive appraisal in future studies. All datasets used originate from previously published work that have undergone thorough ethical review to ensure safe and responsible use (Hofmann et al., 2020; Skerry and Saxe, 2015; Zhan et al., 2023). Any harmful or offensive content was screened and removed. Also, there is no personally identifiable information presented in the data, and all named entities have been masked to ensure privacy.

In this work, several personality traits were reduced to binary for ease of analysis. This leads our findings to a limited coverage of the population and introducing a degree of selection bias commonly observed in real-world scenarios. We emphasise that we do not intend to constrain or imply these per by the limited definitions employed in this paper. While the work presented here facilitated a relatively straightforward estimation of subjectivity, it also raises potential concerns regarding privacy when incorporating personal profile. We believe that future application may benefit from integrating a theory-of-mind perspective (Xu et al., 2025), thereby reducing reliance on explicit personal data.

## Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions. We are grateful to Amy E. Skerry and Prof. Rebecca Saxe for sharing the FGE dataset, and to Ashish Mehta and Kate Petrova for their helpful discussions. This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2), as well as through UKRI Future Leaders fellowship (grant no. MR/Y034295/1 and MR/T041897/1) and the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016).

## References

Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors. 2024. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of ICML*.

AI@Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of COLING*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP*.

Nick Chater, Jian-Qiao Zhu, Jake Spicer, Joakim Sundh, Pablo León-Villagrá, and Adam Sanborn. 2020. Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, 29(5):506–512.

Wei Chen. 2008. Dimensions of subjectivity in natural language. In *Proceedings of ACL*.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of EACL*.

Emma Childs, Tara L White, and Harriet de Wit. 2014. Personality traits modulate emotional and physiological responses to stress. *Behavioural pharmacology*, 25:493–502.

Katherine Maeve Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. 2023. Human uncertainty in concept-based ai systems. In *Proceedings of AIES*.

PT Costa and RR McCrae. 1999. A five-factor theory of personality. *Handbook of Personality: Theory and research*, 2:1999.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron A. Hecht, Jeremy P. Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2:688 – 701.

Wenchao Dong, Assem Zhunis, Dongyoung Jeong, Hyojin Chin, Jiyoung Han, and Meeyoung Cha. 2024. Persona setting pitfall: Persistent outgroup biases in large language models arising from social identity adoption. *arXiv preprint arXiv:2409.03843*.

Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. 2024. Goldcoin: Grounding large language models in privacy laws via contextual integrity theory. In *Proceedings of EMNLP*.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of EMNLP*.

Nicole R Giuliani and James J Gross. 2009. Reappraisal.

Philippe R Goldin, Kateri McRae, Wiveka Ramel, and James J Gross. 2008. The neural bases of emotion regulation: reappraisal and suppression of negative emotion. *Biological psychiatry*, 63(6):577–586.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of HCI*.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

James J Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *Proceedings of ICLR*.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of COLING*.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233.

Leonid V Kantorovich. 1960. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *ICLR*.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Nature Human Behaviour*, 40:e253.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of EMNLP*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Kateri McRae. 2016. Cognitive emotion regulation: A review of theory and scientific findings. *Current Opinion in Behavioral Sciences*, 10:119–124.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.

Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. Uncertain judgements: eliciting experts' probabilities.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*.

Silviu Paun and Edwin Simpson. 2021. Aggregating and learning from multiple annotators. In *Proceedings of EACL*.

Edward J Peacock and Paul TP Wong. 1990. The stress appraisal measure (sam): A multidimensional approach to cognitive appraisal. *Stress medicine*, 6(3):227–236.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP*.

QwenTeam. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Esther Rolf, Nikolay Malkin, Alexandros Graikos, Ana Jojic, Caleb Robinson, and Nebojsa Jojic. 2022. Resolving label uncertainty with implicit posterior models. In *Proceedings of UAI*.

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 and Johnny JR Fontaine. 2013. Driving the emotion process: The appraisal component.

KR Scherer. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Alfred Schütz. 1942. Scheler's theory of intersubjectivity and the general thesis of the alter ego. *Philosophy and Phenomenological Research*, 2(3):323–347.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023a. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023b. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of ACL*.

Mohammad Shokri, Vivek Sharma, Elena Filatova, Shweta Jain, and Sarah Levitan. 2024. Subjectivity detection in english news using large language models. In *Proceedings of WASSA*.

Amy E Skerry and Rebecca Saxe. 2015. Neural representations of emotion are organized around abstract event features. *Current biology*, 25(15):1945–1954.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of CVPR*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of EMNLP*.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Andero Uusberg, Jamie L Taxer, Jennifer Yih, Helen Uusberg, and James J Gross. 2019. Reappraising reappraisal. *Emotion Review*, 11(4):267–282.

Andero Uusberg, Jennifer Yih, Jamie L Taxer, Nicole M Christ, Teili Toms, Helen Uusberg, and James J Gross. 2023. Appraisal shifts during reappraisal. *Emotion*.

Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024a. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. 2024b. Aligning language models with human preferences via a bayesian approach. In *Proceedings of NeurIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of ICLR*.

Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, Caroline Catmur, and Yulan He. 2025. Enigmatom: Improve llms' theory-of-mind reasoning capabilities with neural knowledge base of entity states. *arXiv preprint arXiv:2503.03340*.

Gerard Yeo and Kokil Jaidka. 2023. The peace-reviews dataset: Modeling cognitive appraisals in emotion text analysis. In *Findings of EMNLP*.

Gerard C Yeo and Desmond C Ong. 2024. Associations between cognitive appraisals and emotions: A meta-analytic review. *Psychological Bulletin*, 150:1440–1471.

Hongli Zhan, Desmond C Ong, and Junyi Jessy Li. 2023. Evaluating subjective cognitive appraisals of emotions from large language models. In *Findings of EMNLP*.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models are capable of offering cognitive reappraisal, if guided. In *Proceedings of COLM*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of EMNLP*.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In *Proceedings of EMNLP*.

# A Detail of Appraisal Dimensions

## A.1 Definition of the 21 Appraisal Dimensions

There are various definitions of cognitive appraisal dimensions. In this work, we use the 21 appraisal dimensions defined in (Hofmann et al., 2020). The definition of the 21 appraisal dimensions as listed in Table A1.

## A.2 Unifying Appraisal Dimensions

As the definition of appraisal dimensions vary across EnVent, FGE, and CovidET, we use the 21 appraisal dimensions from EnVent as anchor and manually mapped appraisal dimensions from FGE and CovidET to those of EnVent. The details of appraisal dimension mapping is shown in Table A2.

Further, the appraisal ratings in FGE and CovidET follow a 10-point likert scale instead of the 5-point scale in EnVent. To unify the scale, we normalize the ratings to a 5-point scale by dividing the ratings by 2 and rounding to the nearest integer.

# B Data Examples

We provide examples of data instances from the EnVent, FGE, and CovidET datasets, including the event description and appraisal ratings. We additionally include the demographic information provided in EnVent dataset, which we utilized to carry out experiments §5.1.

**Example from EnVent**

**<Event>**

People get under my skin. Like for example if an entitled customer shows up at my work and demands to speak to my manager for a simple issue that I can resolve. This happens on almost a daily occurrence and it really makes me angry.

**<Appraisal Ratings>**

| Appraisal Dimension | Rating |
|---|---|
| suddenness | 2 |
| familiarity | 5 |
| predict_event | 5 |
| ... | ... |

**<Demographic Information>**

| Info Type | Value |
|---|---|
| previous participation | Yes, first time, I will answer the following questions. |
| age | 18 |
| gender | Male |
| education | High school |
| ethnicity | North American |
| extravert | 2.0 |
| critical | 1.0 |
| ... | ... |

**Example from FGE**

**<Event>**

NAMEVAR was very lost in her Organic Chemistry class, but when she looked at the other students, she seemed to be the only one struggling with the material. She had to raise her hand to ask the professor to explain the problem a second time. The other students in the class chuckled.

**<Appraisal Ratings>**

| Appraisal Dimension | Rating |
|---|---|
| predict_event | 2 |
| pleasantness | 1 |
| goal_support | 5 |
| other_responsibility | 1 |
| ... | ... |

**Example from CovidET**

**<Event>**

I don't even know how to speak of this grief. I have read of many stories of people losing their loved ones, but it didn't happen to my family until today. I lost my uncle who is my best friend, we lived abroad all our lives but because of the pandemic I returned home, and barely managed to see him those 2 years. He was the kindest and purest human being. I met him briefly yesterday as we took him to the hospital. Today he passed away from COVID, he was deathly afraid from needles and the vaccination. I feel so so powerless. So guilty I didn't reply to his phone call 3 days earlier as my family was also sick of Covid and I was caring for them. The death feels like it could have been avoided, he

| Appraisal Dimension | Abbrev. | Description |
|---|---|---|
| suddenness | SD | The event was sudden or abrupt. |
| familiarity | FL | The event was familiar to its experiencer. |
| predict_event | PE | The experiencer could have predicted the occurrence of the event. |
| pleasantness | PL | The event was pleasant for the experiencer. |
| unpleasantness | UPL | The event was unpleasant for the experiencer. |
| goal_relevance | GR | The experiencer expected the event to have important consequences for him/herself. |
| chance_responsibility | CR | The event was caused by chance, special circumstances, or natural forces. |
| self_responsibility | SR | The event was caused by the experiencer's own behavior. |
| other_responsibility | OR | The event was caused by somebody else's behavior. |
| predict_consequence | PC | The experiencer anticipated the consequences of the event. |
| goal_support | GS | The experiencer expected positive consequences for her/himself. |
| urgency | UG | The event required an immediate response. |
| self_control | SC | The experiencer expected positive consequences for her/himself. |
| other_control | OC | Someone other than the experiencer was influencing what was going on. |
| chance_control | CC | The situation was the result of outside influences of which nobody had control. |
| accept_consequence | AC | The experiencer anticipated that he/she could live with the unavoidable consequences of the event. |
| standards | SD | The event clashed with her/his standards and ideals. |
| social_norms | SN | The actions that produced the event violated laws or socially accepted norms. |
| attention | AT | The experiencer had to pay attention to the situation. |
| not_consider | NC | The experiencer wanted to shut the situation out of her/his mind. |
| effort | EF | The situation required her/him a great deal of energy to deal with it. |

Table A1: Definition of appraisal dimensions in (Hofmann et al., 2020)

| | Original | EnVent |
|---|---|---|
| FGE | suddenness | suddenness |
| | familiarity | familiarity |
| | expectedness | predict_event |
| | pleasantness | pleasantness |
| | goal relevance | goal_relevance |
| | agent_intention | chance_responsibility |
| | self_cause | self_responsibility |
| | agent_cause | other_responsibility |
| | certainty | predict_consequence |
| | goal_consistency | goal_support |
| | control | self_control |
| | coping | accept_consequence |
| | self_consistency | standards |
| | moral | social_norms |
| | attention | attention |
| CovidET | familiarity | familiarity |
| | expectedness | predict_event |
| | pleasantness | pleasantness |
| | goal relevance | goal_relevance |
| | self-responsibility | self_responsibility |
| | other-responsibility | other_responsibility |
| | predictability | predict_consequence |
| | goal conduciveness | goal_support |
| | self-controllable | self_control |
| | other-controllable | other_control |
| | circumstances-controllable | chance_control |
| | problem-focused coping | accept_consequence |
| | consistency with internal values | standards |
| | consistency with social norms | social_norms |
| | attentional activity | attention |
| | effort | effort |

Table A2: Mapping of Appraisal Dimensions. *"Original"* indicates the appraisal dimensions in the original dataset, while *"EnVent"* indicates the mapped appraisal dimensions in the EnVent dataset.

took the remaining precautions but it didn't work. At least he died happy alongside his family in his hometown where I live. I don't know what or how to live life without him in my mind, without meeting him ever again, with seeing the same places without him. Im tired, I want to cry. We both hated living at home now I am stuck living on earth without him, I feel this home is the right place for me - was waiting eagerly for him to return. But he barely lasted a week before his death. But mentally I am exhausted of living here, of the pandemic, of having my friends all abroad and getting out of contact, of being alone to face his death. To face this life. Life feels so tasteless.

**<Appraisal Rating>**

| Appraisal Dimension | Rating |
|---|---|
| self_responsibility | 2 |
| other_responsibility | 2 |
| accept_consequence | 1 |
| goal_relevance | 5 |
| attention | 2 |
| ... | ... |

## C Annotation Validity

Each benchmark includes quality-control procedures to ensure annotation validity. We summarize

below the quality assurance protocols reported in the original papers for each datasets.

**EnVent** (Hofmann et al., 2020): Data quality was ensured through multiple steps. Participants were restricted to native English speakers from select countries (US, UK, Australia, New Zealand, Canada, or Ireland) with a Prolific approval rate of $\geq 80\%$. Two types of attention checks were included: one requiring the selection of a specific scale point, and another involving a manual text response. Surveys were restricted to desktop devices to avoid mobile-based auto-corrections. The data collection process was structured in nine rounds, including an initial pilot to gather feedback and adjust instructions, and later rounds to balance underrepresented data. Additional care was taken for dimensions which difficulty for annotators to reliably differentiate.

**FGE** (Skerry and Saxe, 2015): Participants rated how involved the named character was in the narrative. Participants with an average score below 7 were excluded, and individual responses with scores below 5 were discarded. This led to the exclusion of 22 participants, resulting in a final pool of 238 valid subjects and an average of 15.4 responses per stimulus across 200 items. This attention check served as a proxy for engagement and comprehension of the story content.

**CovidET** (Zhan et al., 2023): Includes detailed inter-rater agreement metrics: Krippendorff's alpha (interval) yielded 0.647, indicating substantial agreement; average Spearman's $\rho$ across dimensions was 0.497 (statistically significant); and the mean absolute difference between annotators was low (1.734 on a 1–9 scale). These results indicate that even in a subjective task, annotators exhibited consistent patterns in their judgments.

It is worth noting that CovidET includes limited number of annotators per sample. We incorporate this dataset for two main reasons. First, to evaluate model generalizability on out-of-domain scenario: whereas EnVent and FGE consist of daliy event descriptions, CovidET specifically focuses on situations related to COVID-19, providing a different context for testing the robustness of language models. Second, to examine how the number of annotations per instance affects model performance. We observed that model predictions exhibited similar trends in datasets with a moderate number of annotators, whereas the CovidET dataset showed different model behavior.

## D Analysis of Human Appraisal Ratings

To determine the shape of the distribution for CADE-LSM, we conduct analysis on the mode of distribution of appraisal ratings. As shown in Table A3, regardless of the appraisal dimension, the majority of the appraisal ratings follow some unimodal distribution. In addition, some appraisal ratings display a bimodal distribution. As such, we take two approaches in CADE-LSM where we model the appraisal ratings as a unimodal distribution or as a mixture of two unimodal distributions, effectively reproducing the bimodal distribution.

## E Prompts

Here we list the prompt templates we used to conduct prompting experiments with LLMs. We use the *vanilla* prompt for experiments without auxiliary (demographic) information:

**Vanilla Prompt Template**

**<System Prompt>**
Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? (1 means "Not at all" and 5 means "Extremely")
**</System Prompt>**

**<User Prompt>**
Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? Please rate the situation according to the statements using the Likert scale. The scale ranges from 1 to 5 where 1 means 'Not at all' and 5 means '"Extremely"'. Provide your rating in the following format: "Rating: [Score]". Do not add any explanation or elaboration to your answer.

[Situation]
{{context}}

[Experiencer's Feeling]
{{statements}}
**</User Prompt>**

When conducting experiments that involve auxiliary information such as demographic informa-

| Dimension | *Unimodal* | *Bimodal* | *Trimodal* | *Quadmodal* | *Pentamodal* |
|---|---|---|---|---|---|
| suddenness | 716 | 468 | 0 | 0 | 16 |
| familiarity | 701 | 484 | 0 | 0 | 15 |
| predict_event | 695 | 493 | 0 | 0 | 12 |
| pleasantness | 905 | 293 | 0 | 0 | 2 |
| unpleasantness | 856 | 339 | 0 | 0 | 5 |
| goal_relevance | 725 | 462 | 0 | 0 | 13 |
| chance_responsibility | 813 | 372 | 0 | 0 | 15 |
| self_responsibility | 798 | 394 | 0 | 0 | 8 |
| other_responsibility | 793 | 396 | 0 | 0 | 11 |
| predict_consequence | 686 | 497 | 0 | 0 | 17 |
| goal_support | 786 | 398 | 0 | 0 | 16 |
| urgency | 734 | 446 | 0 | 0 | 20 |
| self_control | 700 | 489 | 0 | 0 | 11 |
| other_control | 734 | 453 | 0 | 0 | 13 |
| chance_control | 784 | 405 | 0 | 0 | 11 |
| accept_consequence | 675 | 506 | 0 | 0 | 19 |
| standards | 843 | 345 | 0 | 0 | 12 |
| social_norms | 941 | 251 | 0 | 0 | 8 |
| attention | 662 | 513 | 0 | 0 | 25 |
| not_consider | 819 | 374 | 0 | 0 | 7 |
| effort | 712 | 476 | 0 | 0 | 12 |

Table A3: Modality of distribution of appraisal ratings in the EnVent dataset. As each appraisal dimension of each situation is rated by five different annotators, we calculate the modality of the distribution ranging from 1 (unimodal) to 5 (uniform).

tion and personality traits, we first flatten the structured information into natural language description. Specifically, for demographic information (as shown in the EnVent data example in § B), we use the demographic information to fill the following template:

```
You are a {age} years old
{ethnicity} {gender} whose
education level is "{education}".
```

For instance, if a given demographic information is

```
Age = 28
Ethnicity = African
Gender = Female
Education = College
```

it will be converted into " `You are a 28 years old African female whose education level is "college".`" We incorporate demographic information into the context by filling the demographic description into the `{{demographic info}}` slot in the following prompt:

**Prompt Template (+Demo)**

**<System Prompt>**
{{demographic info}} Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? (1 means "Not at all" and 5 means "Extremely")
**</System Prompt>**

**<User Prompt>**
{{demographic info}} Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? Please rate the situation according to the statements using the Likert scale. The scale ranges from 1 to 5 where 1 means 'Not at all' and 5 means '"Extremely"'. Provide your rating in the following format: "Rating: [Score]". Do not add any explanation or elaboration to your answer.

[Situation]
{{context}}

[Experiencer's Feeling]
{{statements}}
**</User Prompt>**

Similar to demographic information, we convert the structured personality traits information into natural language descriptions. In the EnVent dataset, the authors used the Big-Five personality traits (Costa and McCrae, 1999). Specifically, the authors leveraged the 10-item assessment of the Big-Five personality traits (Gosling et al., 2003). For each trait of the Big-Five, two descriptions are rated to reflect the annotator's tendency:

```
Openness to experience
                    open
                    conventional

    Conscientiousness
                    dependable
                    disorganized

        Extraversion
                    extraverted
                    quiet

       Agreeableness
                    sympathetic
                    critical

  Emotional Stability
                    calm
                    anxious
```

Each of the 10 items are rated on a scale of 1-7. We compare the rating of the paired items (e.g. comparing `open` and `conventional` for `openness to experience`) and select the item with a higher rating as the description. We omit the personality trait if there is a tie in the rating of the corresponding items. We use the selected item description to fill the following template:

```
You are a {openness}
{conscientiousness}{extraversion}}
{agreeableness}{emotional_stability}
person.
```

We incorporate personality trait information by replacing the `{{personality traits}}` placeholder with the filled personality trait template:

**Prompt Template (+Traits)**

**<System Prompt>**
{{personality traits}} Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? (1 means "Not at all" and 5 means "Extremely")
**</System Prompt>**

**<User Prompt>**
{{personality traits}} Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? Please rate the situation according to the statements using the Likert scale. The scale ranges from 1 to 5 where 1 means 'Not at all' and 5 means '"Extremely"'. Provide your rating in the following format: "Rating: [Score]". Do not add any explanation or elaboration to your answer.

[Situation]
{{context}}

[Experiencer's Feeling]
{{statements}}
**</User Prompt>**

We use the following prompt to incorporate both the demographic information and personality trait information:

**Prompt Template (+Demo +Traits)**

**<System Prompt>**
{{demographic info}} {{personality traits}} Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? (1 means "Not at all" and 5 means "Extremely")
**</System Prompt>**

**<User Prompt>**
{{demographic info}} {{personality traits}}
Put yourself in the shoes of the writer at the time when the event happened, and try to reconstruct how that [Situation] was perceived. How much do these statements apply? Please rate the situation according to the statements using the Likert scale. The scale ranges from 1 to 5 where 1 means 'Not at all' and 5 means '"Extremely"'. Provide your rating in the following format: "Rating: [Score]". Do not add any explanation or elaboration to your answer.

[Situation]
{{context}}

[Experiencer's Feeling]
{{statements}}
**</User Prompt>**

## F Unimodal and Bimodal Label Smoothing

As discussed in §4, in addition to smoothing the appraisal ratings assuming an unimodal shape, we also conducted experiments by assuming that the distribution over ratings follows a bimodal distribution. To simulate a bimodal distribution, we conduct label smoothing using a mixture of two discretized Gaussian distributions, one centered at the ground truth rating and the other at a non-adjacent rating[9]. Experiment results demonstrate that the bimodal CADE-LSM model underperforms that with an unimodal assumption in Wasserstein distance and $\mu$-MAE and achieved minor improvement in $\sigma^2$-MAE. We present the comparison in Table A4

| Model | Wasserstein↓ | $\mu$-MAE↓ | $\sigma^2$-MAE↓ |
|---|---|---|---|
| CADE-LSM-Unimodal | 0.773 | 0.665 | 0.837 |
| CADE-LSM-bimodal | 0.900 | 0.793 | 0.782 |

Table A4: Comparison of unimodal versus bimodal label smoothing results.

## G Analysis of LLMs Appraisal Ratings

A key factor that influences the variance of LLMs' appraisal rating is the `temperature` parameter, which controls the kurtosis of the logits over the vocabulary space. To ensure a fair comparison among LLMs and fine-tuned autoencoding models, we conduct a grid search over temperatures ranging from [0, 1.5]. We provide the Wasserstein distance, $\mu$-MAE, and $\sigma^2$-MAE metrics with respect to the sampling temperature used in Table A5.

| | Temperature | Wasserstein↓ | $\mu$-MAE↓ | $\sigma^2$-MAE↓ |
|---|---|---|---|---|
| Llama3.1-8B | 0.00 | 1.130 | 0.932 | 0.887 |
| | **0.25** | **1.094** | **0.904** | 0.826 |
| | 0.50 | 1.113 | 0.941 | 0.812 |
| | 0.75 | 1.117 | 0.973 | 0.808 |
| | 1.00 | 1.127 | 1.083 | 0.798 |
| | 1.25 | 1.142 | 1.129 | **0.791** |
| Qwen2.5-7B | 0.00 | 1.144 | 0.984 | 0.888 |
| | 0.25 | 1.117 | 0.960 | 0.861 |
| | 0.50 | 1.094 | 0.945 | 0.826 |
| | **0.75** | **1.078** | **0.919** | 0.817 |
| | 1.00 | 1.084 | 0.939 | 0.814 |
| | 1.25 | 1.090 | 0.950 | **0.811** |
| Llama3.3-70B | 0.00 | 1.070 | 0.956 | 0.882 |
| | 0.25 | 1.061 | 0.958 | 0.873 |
| | 0.50 | 1.053 | 0.958 | 0.862 |
| | 0.75 | 1.042 | 0.951 | 0.851 |
| | 1.00 | 1.032 | 0.942 | 0.840 |
| | 1.25 | 1.022 | 0.935 | 0.830 |
| | **1.50** | **1.012** | **0.926** | **0.820** |
| Qwen2.5-72B | 0.00 | 1.092 | 1.021 | 0.870 |
| | 0.25 | 1.056 | 1.027 | 0.830 |
| | 0.50 | 1.022 | 1.025 | 0.799 |
| | 0.75 | 0.995 | 0.945 | 0.771 |
| | 1.00 | 0.974 | 0.892 | 0.749 |
| | 1.25 | 0.954 | 0.873 | 0.738 |
| | **1.50** | **0.945** | **0.852** | **0.736** |

Table A5: Temperature study results.

## H Implementation details

**CADE-VAE:** Given the contextualized representation $\boldsymbol{h} = \text{PLM}(\boldsymbol{s})$ for each situation $\boldsymbol{s}$. We compute the approximation variational posterior $q_\phi(\boldsymbol{z}|\boldsymbol{h})$ using the inference network $\Phi(\boldsymbol{h}; \phi)$:

$$\begin{aligned} \boldsymbol{\mu} &= \boldsymbol{W}_\mu \boldsymbol{h} + \boldsymbol{b}_\mu \\ \log \boldsymbol{\sigma}^2 &= \boldsymbol{W}_\sigma \boldsymbol{h} + \boldsymbol{b}_\sigma \\ \boldsymbol{z} &= \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon} \end{aligned} \tag{3}$$

where $\boldsymbol{W}_\mu$, $\boldsymbol{W}_\sigma$, $\boldsymbol{b}_\mu$, and $\boldsymbol{b}_\sigma$ are parameters for two MLPs. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ define a multivariate Gaussian distribution with a diagonal covariance matrix, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Then, we sample from $q_\phi(\boldsymbol{z}|\boldsymbol{h}) \simeq \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ to generate $\boldsymbol{z} \in \mathbb{R}^l$ as the latent representation, where $l$ is the dimension of the representation. We use inference network

[9]For instance, if the ground truth rating is "2", we would allocate the majority of probability density to ratings "2" and one of ["4", "5"].

| Models | Top Quantified Appraisal Dimensions | | | | |
|---|---|---|---|---|---|
| | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
| CADE-LS | *pleasantness* | *unpleasantness* | *social_norms* | *standards* | *self_responsibility* |
| w. Demo | *pleasantness* | *social_norms* | *unpleasantness* | *standards* | *self_responsibility* |
| w. Traits | *pleasantness* | *unpleasantness* | *social_norms* | *self_responsibility* | *standards* |
| w. Demo & Traits | *pleasantness* | *unpleasantness* | *social_norms* | *self_responsibility* | *standards* |
| Llama3-8B | *social_norms* | *pleasantness* | *unpleasantness* | *self_control* | *chance_responsibility* |
| w. Demo | *social_norms* | *pleasantness* | *unpleasantness* | *self_control* | *Attention* |
| w. Traits | *social_norms* | *pleasantness* | *unpleasantness* | *self_control* | *Attention* |
| w. Demo & Traits | *social_norms* | *pleasantness* | *unpleasantness* | *self_control* | *Attention* |
| Qwen2.5-7B | *pleasantness* | *unpleasantness* | *social_norms* | *self_control* | *Attention* |
| w. Demo | *pleasantness* | *social_norms* | *unpleasantness* | *self_control* | *Attention* |
| w. Traits | *pleasantness* | *social_norms* | *unpleasantness* | *self_control* | *Attention* |
| w. Demo & Traits | *pleasantness* | *social_norms* | *unpleasantness* | *self_control* | *Attention* |

Table A6: Qualitative analysis of subjectivity in well-modeled appraisal dimensions across various models.

| Models | Bottom Quantified Appraisal Dimensions | | | | |
|---|---|---|---|---|---|
| | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
| CADE-LS | *urgency* | *accept_consequence* | *predict_consequence* | *goal_support* | *self_control* |
| w. Demo | *urgency* | *goal_support* | *accept_consequence* | *predict_consequence* | *self_control* |
| w. Traits | *goal_support* | *urgency* | *predict_consequence* | *self_control* | *accept_consequence* |
| w. Demo & Traits | *predict_event* | *urgency* | *predict_consequence* | *accept_consequence* | *self_control* |
| Llama3-8B | *other_control* | *goal_support* | *familiarity* | *predict_consequence* | *chance_control* |
| w. Demo | *other_control* | *goal_support* | *familiarity* | *predict_consequence* | *chance_control* |
| w. Traits | *other_control* | *goal_support* | *familiarity* | *predict_consequence* | *chance_control* |
| w. Demo & Traits | *other_control* | *goal_support* | *familiarity* | *predict_consequence* | *chance_control* |
| Qwen2.5-7B | *familiarity* | *goal_support* | *other_control* | *predict_event* | *accept_consequence* |
| w. Demo | *familiarity* | *goal_support* | *other_control* | *predict_event* | *chance_control* |
| w. Traits | *familiarity* | *goal_support* | *other_control* | *predict_event* | *chance_control* |
| w. Demo & Traits | *familiarity* | *goal_support* | *other_control* | *predict_event* | *chance_control* |

Table A7: Qualitative analysis of subjectivity in poorly modeled appraisal dimensions across various models.

$\Phi(\boldsymbol{h};\phi)$ for inferring $\boldsymbol{z}$ and two-layer parameterized MLP $\Theta(\boldsymbol{h};\theta)$ as the decoder to reconstruct $\boldsymbol{h}$. The parameters can be optimized by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{vae} = \mathbb{E}_{\Phi}[\log \Theta(\boldsymbol{h};\theta)] - \mathrm{KL}(\Phi||p(\boldsymbol{z})) \quad (4)$$

where $p(\boldsymbol{z})$ is the prior follows the Gaussian distribution $\mathcal{N}(0,\mathbf{I})$. We also introduce a regression loss, $\mathcal{L}_{\text{regression}} = \mathrm{MSE}(y,\boldsymbol{h})$ to predict appraisal scores, enhancing text representation for more efficient distribution estimation. The overall objective function is a multi-task learning objective:

$$\mathcal{L} = \mathcal{L}_{vae} + \lambda \mathcal{L}_{\text{regression}} \quad (5)$$

where $\lambda$ is the coefficient that balances the contribution of each component in the training process.

# I Statistical Significance of Personal Profile Incorporation

We have conducted a one-tailed two-sample T-test to analyze the influence of personal profiles to the distribution of ratings produced by LLMs. As shown in Table A8, we found that 1/3 of the appraisal dimensions exhibit statistically significant improvements when either personality traits or demographical information are integrated, across both models. Specifically, adding personality traits improves performance on *familiarity, self_control, urgency, accept_consequence*, and *standards*. Incorporating demographical information yields significant gains on *familiarity*. When both types of personal profiles are integrated, improvements are observed for *predict_event, self_control, accept_consequence*, and *effort*.

For dimensions that significantly improved after adding personality traits, many have strong correlation with certain traits. For instance, *self_control*, which is defined as “the experiencer expects positive consequences for themselves", is highly relevant to traits such as “anxious" and “calm". For dimensions that significantly improved after adding demographic information, the “age", “gender", “education", and “ethnicity" information aid the model to possess a more concrete portrait of the user, hence easier to determine *familiarity*, which represents whether “the event was familiar to its experiencer". LLMs do generally acquire, via pretraining, a strong prior knowledge as to whether a

person from a particular demographic group would be familiar with a given situation. For instance, the event “England scored in the 2nd minute of the Euros final" would be more familiar for people of “European" ethnicity compared to people of “East Asian" ethnicity.

## J Qualitative Study

**Understanding subjectivity from various appraisal dimensions.** Table A6 and A7 present the qualitative study on subjectivity across well-modeled and poorly modeled appraisal dimensions across various models. Figure A1 and A2 show the appraisal distributions predicted by LLMs compared to human annotators for the selected top-ranked and bottom-ranked subjectivity dimensions.

**Effects of Personality Traits** We conduct a qualitative analysis to investigate the effects of various personality traits on the cognitive subjectivity. As shown in Figure A3, certain appraisal dimensions such as *pleasantness* and *unpleasantness* show inherent variance regardless of individuals' personality traits. These variance can largely be attributed to the characteristics of situation descriptions used in EnVent (Hofmann et al., 2020). Specifically, the annotation manual indicates that repeated appraisal ratings are conducted on experiencer-reported situations, which often carry polarized sentiments, as people tend to recall events that evoke strong emotional reactions. I addition, annotators were instructed to describe an event that made them feel one out of twelve predefined emotions, which were deliberately designed to span both postive and negative sentiments. As a results, the reported situations naturally contain bi-polar sentiment, futher amplifying variance in *pleasantness* and *unpleasantness*.

## K Correlations between mean and variance

***Correlations between mean and variance is model-dependent*** To examine whether proficiency in modeling average rating tendency implies competency in capturing subjectivity, we conduct a correlation analysis between the two metrics across various models. We evaluate models using two point-estimate metrics: $\mu$-MAE, which measures a model's ability to capture the population's average rating, and $\sigma^2$-MAE, which quantifies its ability to capture rating subjectivity within the population. Results from Table A9 show that there exists a weak correlation between the two metrics except for the Llama3.1-70B model. Therefore, in addition to making models for representative of the population (e.g. $\mu$-MAE), it is equally important to make them more capable of modeling the nuanced subjectivity within each population (e.g. $\sigma^2$-MAE).

| Dimension | Llama3.1-8B | | | Qwen2.5-7B | | |
|---|---|---|---|---|---|---|
| | w. Demo | w. Traits | w. Demo \& Traits | w. Demo | w. Traits | w. Demo \& Traits |
| *suddenness* | 0.747 | **0.051** | 0.243 | 0.506 | 0.480 | 0.440 |
| *familiarity* | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | 0.700 |
| *predict_event* | **0.001** | 0.236 | **0.008** | 0.515 | **0.086** | **0.067** |
| *pleasantness* | 0.372 | 0.885 | 0.936 | 0.616 | 0.351 | 0.825 |
| *unpleasantness* | 0.936 | 0.948 | 0.276 | 0.598 | 0.597 | 0.633 |
| *goal_relevance* | 0.258 | **0.000** | 0.973 | **0.093** | 0.621 | **0.008** |
| *chance_responsblt* | 0.419 | 0.756 | 0.187 | **0.070** | **0.034** | 0.424 |
| *self_responsblt* | **0.047** | 0.100 | **0.020** | 0.273 | 0.135 | 0.122 |
| *other_responsblt* | 0.183 | 0.632 | 0.792 | 0.452 | 0.308 | **0.046** |
| *predict_conseq* | **0.010** | **0.000** | **0.034** | 0.909 | 0.173 | 0.754 |
| *self_control* | **0.001** | **0.000** | **0.014** | 0.114 | **0.000** | **0.014** |
| *urgency* | 0.752 | **0.015** | 0.862 | 0.273 | **0.016** | **0.034** |
| *other_control* | **0.000** | 0.435 | 0.105 | 0.965 | 0.273 | **0.019** |
| *chance_control* | 0.115 | 0.101 | 0.069 | **0.100** | **0.008** | **0.001** |
| *accept_conseq* | **0.015** | **0.022** | **0.042** | 0.264 | **0.015** | **0.000** |
| *standards* | 0.866 | **0.000** | 0.772 | 0.668 | **0.062** | 0.104 |
| *social_norms* | 0.631 | 0.847 | 0.979 | 0.146 | **0.030** | **0.000** |
| *attention* | **0.033** | **0.005** | 0.755 | 0.702 | 0.498 | 0.357 |
| *not_consider* | 0.272 | **0.000** | 0.952 | 0.637 | 0.752 | 0.269 |
| *effort* | 0.576 | **0.001** | **0.017** | 0.641 | 0.610 | **0.000** |
| *goal_support* | 0.983 | 0.548 | 0.530 | 0.151 | 0.421 | 0.516 |

Table A8: Statistical significance of personal profile incorporation

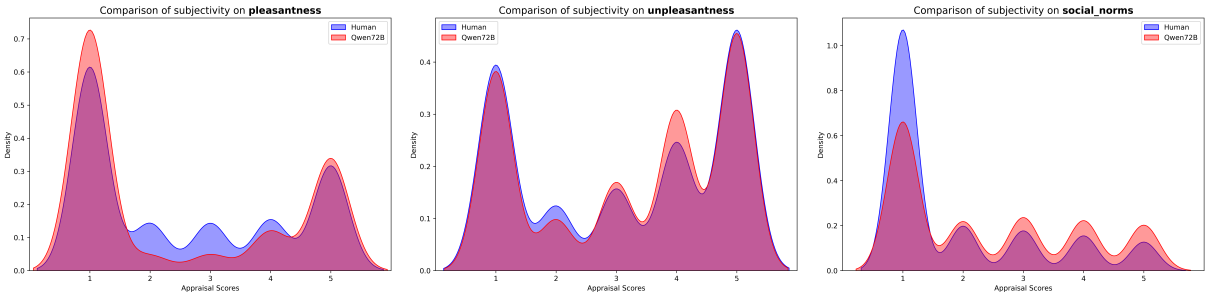


Figure A1: Comparison of appraisal distributions for top-ranked subjectivity dimensions between human annotators and Qwen-72B.

| Model | Metric1 ($m1$) | Metric2 ($m2$) | $\rho_{m1,m2}$ |
|---|---|---|---|
| Llama3.3-8B | $\sigma^2$-MAE | $\mu$-MAE | 0.596 |
| Llama3.3-8B | $\sigma^2$-MAE | Wasserstein | 0.683 |
| Qwen2.5-7B | $\sigma^2$-MAE | $\mu$-MAE | 0.292 |
| Qwen2.5-7B | $\sigma^2$-MAE | Wasserstein | 0.538 |
| Llama3.1-70B | $\sigma^2$-MAE | $\mu$-MAE | 0.832 |
| Llama3.1-70B | $\sigma^2$-MAE | Wasserstein | 0.922 |
| Qwen2.5-72B | $\sigma^2$-MAE | $\mu$-MAE | 0.369 |
| Qwen2.5-72B | $\sigma^2$-MAE | Wasserstein | 0.544 |

Table A9: Pearson correlation between the two point-estimate metrics across 4 models.

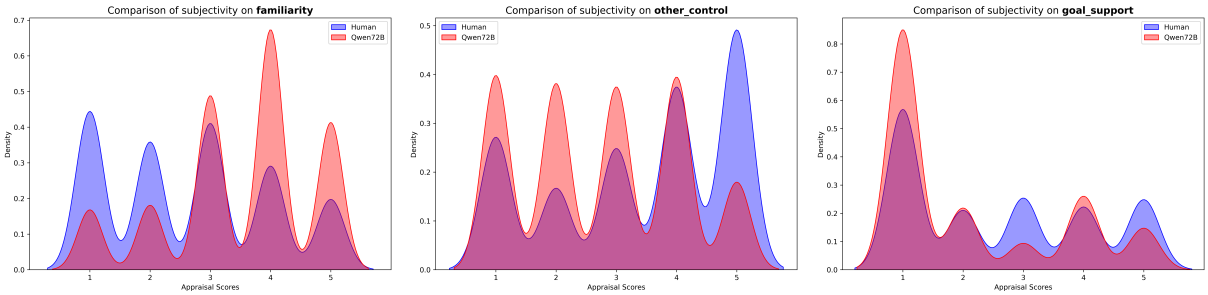


Figure A2: Comparison of appraisal distributions for bottom-ranked subjectivity dimensions between human annotators and Qwen-72B.

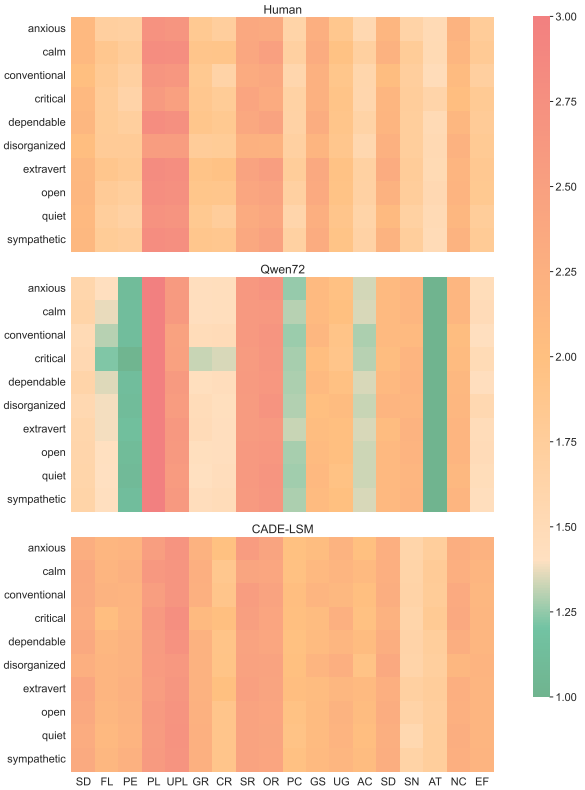


Figure A3: Comparison of appraisal variance in different personalities in the EnVent dataset.