Secrets of a Smile: Can Internal Experiences be Derived from External Expressions?

Mirella Hladký[*†], Jozef Hladký[‡], Tanja Schneeberger[†], Patrick Gebhard[†], Tomohiro Nagashima[*], Philipp Müller[§]
[*]*Saarland University*, *Saarland Informatics Campus, Saarbrücken, Germany*, firstname.lastname@uni-saarland.de
[†]*German Research Center for Artificial Intelligence*, *Saarbrücken, Germany*, firstname.lastname@dfki.de
[‡]*Independent Researcher*, *Zürich, Switzerland*, firstname@lastname.de
[§]*Max Planck Institute for Intelligent Systems*, *Stuttgart, Germany*, mueller@is.mpg.de

***Abstract*—Computational emotion recognition traditionally relies on observable signals, but this approach fails in emotionally challenging social situations, where individuals actively regulate their expressions, e.g., through smiles and carefully chosen words. This regulation creates a critical mismatch between internal experience and external expressions, leading systems that typically associate positive expressions with positive emotions to misinterpret these signals. In this paper, we study to what extent self-reported internal functions of smiles can be decoded from multimodal observable behavior. To this end, we extend an existing corpus of smiles in shame-eliciting situations by adding comprehensive verbal annotations, creating a unique multimodal resource of 172 smile instances from 27 participants that enables comprehensive analysis of social and emotional signals in challenging contexts. Analyses using both computational models and human observers failed to reliably interpret smiles, e.g., whether a smile reflects or masks an internal emotion or serves self-regulatory, relationship-maintaining, or relationship-compromising functions. Our results suggest fundamental limitations of interpreting internal states based purely on observable expressions, and that a reliable interpretation of complex internal states may require intrapersonal information from introspective self-reports. This work contributes novel multimodal annotations, available as an open research resource, and provides evidence that between-person variability limits universal approaches to interpreting smile functions.**

***Index Terms*—Emotion Recognition, Emotion Regulation Multimodal Corpus, Facial Coding, Verbal Analysis**

I. Introduction

In a job interview, an applicant smiles when receiving negative feedback – while smiles are typically associated with positive emotions, the applicant self-reports intense negative emotions and self-protective regulation [1]–[3]. This paradox exposes a critical flaw in computational emotion recognition: the implicit assumption that expressions directly reflect internal states. In negative social situations, individuals actively regulate both verbal and nonverbal channels to protect self-concept and relationships [4]–[7]. This creates mismatches between internal experiences and external expressions [8] that mislead systems into interpreting protective displays as genuine affect. We bridge this gap by modeling the interplay between regulated verbal and nonverbal signals to predict their subjective functions. Our approach captures smile-related verbal content and communicative intention alongside nonverbal smile features. We investigate whether internal functions of smile expressions can be predicted based on multimodal observable expressions in the context of negative social situations. Using self-reported smile functions as ground truth, we apply three different approaches to answer this question: 1) machine learning classification, 2) human annotation, and 3) Bayesian regression modeling. The contributions of our work are threefold:

- Novel lexical and functional verbal annotations from manual transcripts on a corpus of 172 smile instances, provided as an open research resource.
- A systematic evaluation of smile function prediction using machine learning, human annotation, and Bayesian regression modeling.
- Evidence that individual differences create substantial and irreducible variance, pointing to fundamental limits of universal decoding approaches.

II. Background and Related Work

A. Emotion Regulation and the Approach-Distance Dilemma

Emotions arise in situations relevant to one's goals, involving experiential, behavioral, and physiological components [6]. We distinguish *internal experiences* of emotion, which are not directly observable, from *external expressions* that manifest in verbal and nonverbal signals [9], [10]. Critically, these two dimensions do not map directly; the emotional process involves an appraisal of a stimulus, the generation of an internal experience, and the subsequent external expression [11]. This process can be modified, either consciously or unconsciously, through *intrapersonal emotion regulation strategies*, influencing experienced and expressed emotions [4], [6], [12], while *interpersonal emotion regulation* aims to influence others' emotions through specific verbal and nonverbal expressions [13]. Consequently, observable signals can encode various regulatory processes: an individual may smile to mask their internal state, to influence their social standing, or improve their own well-being [1], [14].

Such regulatory complexity becomes particularly evident in high-stakes evaluative contexts such as job interviews [15]. When an applicant feels that they do not meet expectations, social emotions like shame become particularly salient [2], [3] – a self-conscious emotion resulting from negative self-

979-8-3195-2145-3/26/$31.00 ©2026 IEEE

evaluation that threatens the self-concept (an individual's beliefs about themselves [16]) [5]. Shame can also arise when an individual fears rejection from someone whose opinion they regard as important [17], [18]. A shame-eliciting situation activates strategies to defend the self-concept [19], [20] while simultaneously avoiding social rejection [5], [12], [18]. This creates an *approach-distance dilemma* [21], [22]: the applicant may blame themselves and give an appeasing smile (approach), or express critique and withdraw [7], [14] (distance). The dilemma's resolution depends on individual differences in emotion regulation strategies and interaction goals [4], [23], motivating individuals toward relationship maintenance (approach) or self-protective distancing [13], [21], [22].

B. Functions of Expressions in Negative Social Situations

Both humans and computational emotion recognition rely on expressed verbal and nonverbal signals and their coordination to interpret internal states [24]–[26]. However, in social interactions shaped by an approach-distance dilemma, verbal and nonverbal signals often convey incongruent emotional information, limiting their interpretability [24]. Critically, in such situations, both nonverbal and verbal expressions serve emotion and relationship regulating functions rather than directly reflecting internal states. Especially smiles are often used to manage emotionally charged social situations [4], [7], [27], [28]. Expressions in negative social situations can be largely categorized along two dimensions: (1) *interpersonal vs. intrapersonal* and (2) *relationship approaching vs. distancing*.

a) Nonverbal signals: Smiles are particularly multifunctional [29], [30]. Beyond expressing happiness, they regulate emotions intrapersonally [4], [7], [27], [28] and interpersonally, managing relationships [7], [30], [31]. Their morphological features, including Duchenne vs. Non-Duchenne activation, Symmetry, and Smile Controls (e.g., pressed lips, depressed lip corners) [31]–[33] can hint at specific functions [1], [30], [34], [35]. Hladký et al. [1] developed a function-based taxonomy of smiles from self-reports in negative social contexts, identifying three primary categories: *Representative functions* (smiles representing negative internal states), frequently appearing as controlled Duchenne smiles; *Intrapersonal functions* (reappraisal-based self-regulation), enhancing well-being through intense or Duchenne smiles [1], consistent with research showing smiles can induce positive affect [27], [28]; *Interpersonal functions* encompass three sub-functions: *relationship supporting* functions (intense, Non-Duchenne smiles), *relationship adjusting/compromising* functions of signaling dominance, conflict, or devaluation (asymmetric and controlled smiles, related to known contempt [31] and dominance displays [34]), and *emotion masking* functions (associated with face-touching) [1]. Critically, the mapping between morphological features and functions is rather ambiguous: smiles are shaped by individual characteristics such as habitual emotion regulation strategies and social values [23], variables that are not directly observable and must be self-disclosed.

b) Verbal signals: In naturalistic settings where nonverbal signals alone are ambiguous, verbal expressions may provide additional information, as the preceding, accompanying, and subsequent verbal context of a smile shapes its meaning [36]. Verbal signals regulate negative emotions [37], [38] and social relationships [39]. At the *lexical level* (e.g., analyzed using *LIWC* [40]), first-person pronouns indicate self-directed attention, negative emotion and lower social status [41], [42]; second-person pronouns indicate other-directed attention but also conflict [43], [44]; tentative language signals uncertainty [45]; emotion words directly capture affective expression and motivational drives capture status concerns relevant in shame-eliciting contexts [7], [20]. At the *functional level*, *Dialogue Acts (DAs)* capture communicative intentions (e.g., apologize, criticize, agree) [46], [47], revealing speakers' social intentions and interpersonal emotion regulation strategies [48]–[51]. In approach-distance dilemmas, DAs can signal approach (e.g., agree, apologize) or distancing (e.g., disagree, criticize, blame) intentions. Verbal fillers (e.g., 'umm') can indicate uncertainty and ongoing intrapersonal regulation processes that can temporarily stall the conversation.

C. Inferring Internal States from External Expressions

Because regulatory processes obscure internal emotional experiences, traditional approaches assuming a direct mapping between external expressions and internal states are unsuitable for negative social situations [8]. The integration of nonverbal and verbal channels improves emotion recognition only when both channels convey congruent information [24]. However, in high-stakes social situations, they often carry incongruent content, due to display rules, impression management, and social intentions [52]–[54]. Modern psychology challenges the direct-mapping assumption: mappings between facial expressions and felt emotions are modest, and expressions cannot be decoded as a straightforward readout of internal experience [55]. Affective computing has begun addressing this complexity through context-aware recognition methods [56], [57] and regulation-aware models (e.g., MARSSI [58], DEEP [59]) which demonstrate that emotion regulation must be considered when interpreting expressions. Hladký et al. [1] linked smile morphology to self-reported functions in negative social situations, showing that smiles serve diverse intra- and interpersonal regulatory purposes. However, their classification did not clearly outperform trivial baselines, was limited to nonverbal signals, and did not account for individual differences. This prior work failed to reliably decode internal smile functions, leaving open whether this reflects insufficient signal coverage, methodological limitations, or a fundamental ambiguity in the mapping from observable expressions to internal states. To distinguish between methodological limitations and fundamental ambiguity, we (1) extend the feature space to include verbal signals, (2) establish a human performance ceiling, and (3) use regression modeling that quantifies between-person variability.

III. Smile Corpus

This work builds upon the empirically grounded smile corpus presented in [1]. It investigated the relationship between smile morphology and internal functions in negative social situations. The corpus contains multi-modal data from $N = 30$ German participants (24 female; 6 male; age 19-58, $M = 30.33$, $SD = 10.59$) who participated in an online job interview led by a virtual agent. The interview used a validated shame-elicitation protocol [3], [59] that successfully induced shame. Participants were confronted with three shame-eliciting statements in the beginning, middle and end of the job interview, targeting their physical appearance, professional experience, and interview performance (details in [1], [3], full job interview script on OSF[1]).

The extended corpus introduced in our work adds a new layer of annotated verbal data to the existing nonverbal and self-report data. To capture smile-specific verbal dynamics [36], we defined relative duration thresholds to non-exclusively classify verbal annotations as occurring before (*pre-smile*), during (*in-smile*) or after a smile (*post-smile*) relative to each smile instance (classification algorithm and source code on OSF[1]). We provide all corpus annotations as an *open research resource* on OSF[1] including an *interactive visualizer tool* to view temporal alignments of nonverbal smile signals and accompanying verbal signals.

A. Smile Corpus Data

We use existing smile corpus data by [1] on the two variables: *smile functions* and *smile-related nonverbal signals*, and extend it by adding a new variable *smile-related verbal signals*. We also include the *situational context* variable from the original corpus.

*1) **Smile Functions**:* Hladký et al. [1] obtained ground truth labels for smile functions via semi-structured post-interviews, where participants reviewed video recordings of themselves in shame-eliciting situations and reported on their internal experiences. A total of 199 smile instances were identified, with participants assigning one or multiple functions to each instance (a total of 364 functions). Of 199 smile instances, we use 172 as we had to exclude 27 due to missing audio or poor quality. Qualitative content analysis of these self-reports resulted in a category system comprising three main smile function types [1]:

Representative Functions encompass smiles that outwardly express an experienced emotional state. *Intrapersonal Functions* comprise smiles that indicate and serve private self-regulation, such as modulating one's own emotional state. The third broad class, *Interpersonal Functions*, groups smiles that manage social relationships; it contains three subclasses: *Relationship Supporting* (smiles that improve or preserve relationships), *Relationship Adjusting* (smiles that signal distancing, devaluation, or a recalibration of social status), and *Mask Internal Emotions* (smiles that hide the smiler's authentic affect behind a facial display). Details of this smile taxonomy can be found in [1]. We use these categories (prevalence in parentheses) in our analyses: Representative (33.9%), Intrapersonal (31.0%), Relationship Supporting (40.9%), Relationship Adjusting (36.8%), Masking Emotions (19.3%).

[1] https://osf.io/dntjv/overview?view_only=74973ef40f484b338a2cc17267d44d83

*2) **Smile-related Nonverbal Signals**:* The nonverbal annotations in the original smile corpus [1] capture facial actions associated with a smile: *Duchenne* markers (eye-muscle contraction alongside raised lip corners) versus *Non-Duchenne* (lip corners raised without eye involvement); *Non-Symmetric* (one-sided smile); *Intensity* (teeth visible); *Laughter* (audible vocalization); *Smile Control* (actions that suppress or alter a smile, e.g., pressed or pursed lips, depressed lip corners, raised chin); *Gaze Aversion*; and *Adaptors* (face or body self-touch). See [1] for the full coding manual.

*3) **Smile-related Verbal Signals**:* We extend the smile corpus by annotating verbal expressions accompanying smiles using quantitative and qualitative text analysis methods:

*a) **Quantitative Lexical Verbal Analysis (LIWC)**:* We analyzed word-level features using the German adaptation of Linguistic Inquiry and Word Count (LIWC) [40], a standard tool for quantifying psycholinguistic categories. We selected six categories based on their theoretical relevance to the shame-eliciting context (see Sec. II-B) while maintaining a manageable predictor-to-observation ratio: *first-person pronouns*, *formal second-person pronouns*, *tentative language*, *positive* and *negative emotion words*, and *motivational drives*.

*b) **Qualitative Functional Verbal Analysis (Dialogue Acts)**:* To capture communicative functions we employed Dialogue Act (DA) analysis, following Bunt et al.'s ISO standard [60]. We categorize DAs as reactions to the approach-distance dilemma (see Sec. II-A) present in our data into two main categories: (1) *Approach* – DAs that carry intentions to maintain/enhance the relationship to the interviewer (e.g., Self-Present, Justify, Agree, Acknowledge, Apologize, Thank, Blame Self); (2) *Distance* – DAs that carry intentions to compromise/abandon the relationship (e.g., Disagree, Challenge, Withdraw, Criticize, Disengage). We include general-purpose communicative functions (e.g. Agree, Disagree, Apologize) presented in [60] and introduce domain-specific functions [49] (e.g., Blame Self, Disengage, Self-Present) to capture emotion and relationship regulating functions in negative social situations. Our smile function predictors include *Approach* (N = 113), *Distance* (N = 60) and a time-management category *Stall* (verbal fillers; N = 158). The full DA annotation scheme can be found on OSF[1].

We employed a human-led deductive coding approach [61]. First, the human coder developed a theoretically grounded annotation scheme and validated it on sample transcripts, iteratively refining category definitions and segmentation rules. Next, to improve efficiency, we used an LLM for pre-coding. We translated the validated scheme into a structured system prompt (following recommendations by [62]) and conducted initial LLM trial run-throughs, with the human coder identifying discrepancies between LLM outputs and theoretical framework to refine the prompt [63], [64]. The resulting system prompt can be found on OSF[1]. In a final step, the human coder reviewed every automated output, corrected inappropriate classifications, and made all final coding decisions [63], [65]. Following best practices for computational assistance in qualitative analysis [62], [66], we configured the open-source

Kimi K2.5 model (Mixture-of-Experts architecture) with low-temperature parameters (temperature = 0.1, top-p = 0.4) and chain-of-thought reasoning to generate transparent classifications [62], [66]. LLM traffic was end-to-end encrypted and ran in a protected enclave, avoiding third-party sharing [67].

*4) **Situational Context**:* A label Situation defines start and end of three video sections that are considered for analysis containing participants' responses in the three shame-eliciting situations. We use this variable to capture context effects, as they contain three different shame-eliciting statements occurring at the beginning, middle and end of the job interview (details in [1], [3], full job interview script on OSF[1]).

IV. Smile Function Prediction

Using the corpus described in Sec. III, we address our Research Question: *Can internal functions of smile expressions be predicted based on observed nonverbal and verbal signals in the context of negative social situations characterized by an approach-distance dilemma?*

We applied three different approaches: 1) machine learning classification and 2) human annotations to assess whether smile functions can be predicted from observable nonverbal and verbal signals; and 3) statistical regression modeling to identify which specific predictors are reliably associated with each function while accounting for between-person variability. All approaches predicted five smile functions as binary outcomes: Representative, Intrapersonal, Relationship-Supporting, Relationship-Adjusting, and Masking Emotions. Predictors for machine learning and regression included seven Nonverbal Signals, three Dialogue Act categories, six Linguistic categories and Situational Context (see Section III).

A. Machine Learning Classification

To classify smile functions, we utilize Support Vector Machines (SVMs) with radial basis function (RBF) kernels. In line with previous work [1], we train separate classifiers for each of the five smile function ground truth classes. Training and evaluation proceeds via leave-one-subject-out nested cross-validation. The outer cross-validation loop is used to evaluate classifier performance on an unseen test subject. The inner loop is used to find optimal hyperparameters. The regularization parameter C is chosen from $\{0.01, 0.1, 1, 10, 100\}$, and the kernel parameter γ is chosen from $\{scale, 0.001, 0.01, 0.1, 1\}$, where $scale$ denotes the default RBF kernel parameter in scikit-learn [68], i.e. the inverse of the number of features times input variance.

To evaluate classifier performance under unbalanced smile function class distributions we report balanced accuracy as the primary metric as it is not inflated by majority-class predictions. We additionally report precision, recall, and F1, with chance F1 (stratified-random baseline) and chance accuracy (majority-class baseline) contextualizing performance relative to the varying class prevalence. The metrics enable direct comparison with human rater performance (see Section IV-B).

Results: All classifiers failed to perform above the majority-class baseline on balanced accuracy across all five categories (see Table I). The SVM trained on all features achieved a macro-averaged balanced accuracy of .538 and macro F1 of .321, both below trivial baselines. While some individual categories showed F1 scores above stratified-random chance F1, balanced accuracy remained more informative: for Mask Emotions, the above-chance F1 reflects that its low prevalence lowers chance F1, while for Relationship Adjusting, high recall and low precision indicate many false positives. Balanced accuracy remained below the majority-class baseline for every category. Notably, the classifier mostly failed to detect Intrapersonal smiles (recall = .038), predicting only 2 of 53 positive cases. This finding is in line with [1], where classifiers likewise failed to consistently outperform the majority-class baseline when class imbalance was accounted for.

B. Human Annotation

Internal smile functions were rated by four trained German raters (2 male, 2 female, age 26-54, *M* = 35.75) who independently annotated the 172 smile instances. We evaluated the raters' ability to perceive and understand own and others' mental states with the Mentalizing Emotions Questionnaire (MEQ) [69]. Sixteen items assess their ability with regards to self (e.g., *I try to understand the different reasons for my feelings*), communication (e.g., *I can explain my different feelings to others*), and others (e.g., *I can perceive conflicting feelings in others*) on a 7-point frequency scale (1 = *never*, 7 = *always*). Overall (possible score range 16-112), they had a score of *M* = 96.25 (*SD* = 9.78) and in the subscale reflecting mentalizing abilities with regards to others (possible range 7-49), *M* = 41.25 (*SD* = 3.95), both above normative average [69]. All raters received identical instruction on the five function categories (Representative, Intrapersonal, Relationship Supporting, Relationship Adjusting, Mask Emotions) and were familiarized with the full job interview context, including the shame-elicitation protocol. For each of the 81 shame-eliciting situations, raters first watched the complete interaction sequence, then viewed each smile instance individually, assigning one or more functions per instance to account for multi-label occurrences in the ground truth data. Raters were free to re-watch as needed.

Agreement across all four raters, measured with Krippendorff's α (robust to imbalanced categories and small rater counts), was low [70]. The macro-averaged α across categories was .062, indicating *unreliable agreement* ($\alpha < .667$). *Relationship Supporting* ($\alpha = .186$) and *Adjusting* ($\alpha = .143$) had the highest (although still poor) agreement, *Representative* was at chance level ($\alpha = .025$) while negative values for *Intrapersonal* ($\alpha = -.030$) and *Mask Emotions* ($\alpha = -.016$) indicate a tendency for systematic disagreement among raters [71].

Results: Human raters failed to perform beyond chance level in assigning smile functions (as self-reported in ground truth data) to smiles. Table I reports precision, recall, F1, balanced accuracy, and chance baselines for human raters (mean across four raters) in comparison to the SVM classifier. Both human raters and the SVM classifier performed comparably, with near-identical macro-averaged balanced accuracy. Social-

TABLE I: Classification metrics for human raters and SVM classifier.

Category	Precision	Recall	F1	Chance F1	Balanced Accuracy	Chance Accuracy
Human Raters						
Representative	.396 (.070)	.427 (.222)	.378 (.077)	.339	.529 (.038)	.661
Intrapersonal	.306 (.040)	.283 (.208)	.265 (.114)	.310	.506 (.017)	.690
Rel. Supporting	.536 (.066)	.286 (.105)	.357 (.074)	.409	.550 (.013)	.591
Rel. Adjusting	.507 (.043)	.365 (.088)	.418 (.057)	.368	.577 (.025)	.632
Mask Emotions	.227 (.028)	.614 (.186)	.321 (.020)	.193	.541 (.038)	.807
Macro-avg.	.394 (.016)	.395 (.156)	.348 (.058)	.324	.540 (.008)	.676
SVM (all features)						
Representative	.333	.328	.330	.339	.496	.661
Intrapersonal	.500	.038	.070	.310	.510	.690
Rel. Supporting	.400	.257	.313	.409	.495	.591
Rel. Adjusting	.450	.714	.552	.368	.603	.632
Mask Emotions	.316	.364	.338	.193	.588	.807
Macro-avg.	.400	.340	.321	.324	.538	.676

Note. Human raters: *Mean (SD)* across four raters. Chance Acc. = majority-class baseline. Chance F1 = expected F1 of a stratified random classifier.

communicative functions (Relationship Adjusting and Supporting) were identified relatively more reliably than intrapersonal regulatory functions and masking functions, consistent with the inter-rater pattern.

C. *Statistical Regression Modeling*

Both classification and human annotation indicate that observable signals do not carry sufficient information for predicting internal smile functions. However, neither approach reveals *which* signals, if any, are reliably associated with specific functions, nor how much unpredictability is due to differences between individuals. We investigate these questions with confirmatory regression modeling, which estimates directional associations of individual predictors while explicitly accounting for between-person variability via random intercepts.

Smile functions were modeled as binary outcomes using Bayesian Generalized Linear Mixed Models (GLMMs) with participant-specific random intercepts. The GLMM framework addresses the data structure by: (1) explicitly modeling between-person variability in baseline smile tendencies via random intercepts; (2) jointly estimating all five smile functions while accounting for their within-person correlations via a multivariate normal structure; and (3) providing uncertainty quantification through Bayesian inference. Situational context was modeled as a categorical fixed effect using orthonormal contrasts (Contrast 1 = Situation 1 vs. Situation 2; Contrast 2 = Situations 1 and 2 combined vs. Situation 3). Priors were assumed to be weakly informative: Normal(0, 1.5) for fixed effects, HalfNormal(1) for random intercept SDs, regularizing against overfitting without strongly constraining effect directions in cases of sparse or unbalanced data. Convergence was confirmed by R-hat < 1.01 and effective sample sizes > 2400 across all parameters.

Results: Random intercept standard deviations were substantial across outcomes (σ_u: 1.04–1.81), indicating considerable between-person variability in baseline smile tendencies. Fixed effects explained 19–37% of latent variance (marginal R^2), while the full model including person-level variability explained 40–68% (conditional R^2). The random intercept structure showed a credible correlation between Intrapersonal and MaskEmotions ($r = .51$, 95% CI .17, .83).

Table II reports fixed-effect coefficients for predictors that showed credible effects on at least one smile function. *Representative* smiles were predicted by Gaze Aversion, Laughter, and Non-Duchenne smiling, negatively by Non-Symmetric smiles. *Intrapersonal* smiles were predicted by smile Intensity and negatively by Duchenne smiling; distance-oriented DAs and motivational drives language were positive predictors while first-person pronouns were negative. *Relationship-Adjusting* smiles were predicted by Non-Symmetric smiles and first-person pronouns, negatively by Gaze Aversion. Smiles that *Mask Emotions* were predicted positively by Non-Duchenne and negatively by Non-Symmetric smiles and language indicating motivational drives. *Relationship-Supporting* smiles showed a positive effect only for first-person pronouns. Regarding situational context, Intrapersonal and Masking smiles were most likely in Situation 1, while Relationship-Adjusting smiles were most likely in Situation 3. An additional exploratory analysis split dialogue acts by timing (pre-, during-, post-smile windows) and revealed that distancing DAs occurring specifically during smiling ($b = 0.90$, 95% CI [0.02, 1.80], OR $= 2.45$) were more strongly associated with Intrapersonal smiles than when aggregated across windows. Full posterior summaries (b, 95% HDI, odds ratios) for all credible effects are provided on OSF[1].

While the preceding results revealed only scattered and function-specific predictor effects, the random intercept standard deviations suggest substantial *between-person variability*. To explore the magnitude of this variability, we computed *intraclass correlation coefficients (ICCs)* that quantify the proportion of total variance in smile function outcomes attributable to between-person differences. ICCs $> .05$ are considered small, $> .20$ medium, and $> .40$ large [72]. Unconditional ICCs (intercept-only models) ranged from .12 to .26 (mean $= .18$), indicating meaningful between-person variation prior to accounting for any predictors. Conditional ICCs (full model

TABLE II: Fixed effects from the multivariate Bayesian GLMM predicting smile functions.

		Representative	Intrapersonal	Relationship Supporting	Relationship Adjusting	Masking Emotions
Nonverbal	Duchenne	−0.38	**−1.24***	0.43	0.76	0.14
	Non-Duchenne	**0.87***	−0.04	0.49	−0.57	**1.18***
	Non-Symmetric	**−0.85***	−0.07	0.19	**1.17***	**−1.20***
	Intensity	−0.56	**1.26***	−0.42	−0.12	0.78
	Laughter	**1.30***	0.47	−0.28	−0.76	−1.24
	Gaze Aversion	**0.28***	0.05	0.11	**−0.43***	−0.00
Dialogue Act	Distance	0.25	**0.72***	−0.20	0.02	−0.02
LIWC	First-person ("I")	−0.16	**−0.44***	**0.47***	**0.51***	0.42
	Motivational Drives	0.22	**0.43***	−0.37	−0.48	**−0.65***
Situation	Contrast 1 (1 vs. 2)	−0.01	**1.03***	0.04	**−1.32***	**1.75***
	Contrast 2 (1&2 vs. 3)	0.47	0.20	0.20	**−1.22***	−1.00

Note. Coefficients (b) are log-odds on the standardized scale. **Bold*** = 95% HDI excludes zero (credible effect). The full model includes 18 predictors; predictors without credible effects for any outcome are omitted. See supplementary table on OSF[1] for all coefficients.

including all predictors) ranged from .21 to .46 (mean = .32), as shown in Table III. Notably, conditional ICCs were larger than unconditional ICCs for all five smile functions. Even when including nonverbal, verbal, and situational context, a substantial portion of variance in smile function remains attributable to the individual displaying the smile.

TABLE III: Intraclass Correlation Coefficients (ICCs) for Smile Functions.

Smile Function	Unconditional ICC	Conditional ICC	% Between-Person[a]
Representative	.12 [.00, .37]	.26 [.02, .57]	25.8%
Intrapersonal	.16 [.00, .41]	.32 [.05, .61]	32.1%
Rel. Support	.14 [.00, .35]	.21 [.00, .52]	21.3%
Rel. Adjustment	.26 [.05, .51]	.35 [.06, .63]	34.9%
Mask Emotions	.22 [.01, .49]	.46 [.14, .69]	45.5%
Mean	.18	.32	31.9%

Note. Brackets indicate 95% highest-density credible intervals. Unconditional ICCs: intercept-only models. Conditional ICCs: full model including nonverbal, verbal, and situational predictors.
[a]Percentage of total latent variance attributable to between-person differences in the conditional model.

V. Discussion

This study investigated whether internal smile functions can be inferred from observable cues in shame-eliciting interactions – probing the fundamental gap between internal emotional experience and external expression [9]–[11]. Using both machine learning classification and regression modeling, we consistently found that observable nonverbal and verbal signals, as well as situational context, provide only limited information about internal smile functions. This conclusion is reinforced by human annotation performance: trained raters with full contextual knowledge could not agree with each other or with ground truth beyond chance level. We identified function-specific associations but no single signal or signal combination reliably discriminating across smile functions.

Previous work [1] investigating the link between smile morphology and smile functions in negative social situations did not show a clear mapping, and focused exclusively on nonverbal signals without verbal and situational context or accounting for differences between individuals. Our results replicate and extend previous findings in four ways. *First*, our regression model confirmed associations between specific nonverbal signals and individual smile functions. Laughter, Non-Duchenne smiles and smile symmetry represent negative emotions, contradicting previous findings [1]. Symmetric Non-Duchenne smiles are generally considered to serve social functions by masking negative emotions [31] rather than expressing them. Intense smiles indicate intrapersonal regulation, in line with [1], [27]; non-symmetric smiles and maintained gaze indicate relationship adjusting/compromising intentions, as in [1], similar to 'false' [31] and 'dominance' smiles [34] and reflecting cues of other-attacking shame regulation [19]; symmetric Non-Duchenne smiles mask emotions, a result not found by [1] but supporting traditional assumptions [31]. We found no nonverbal signal linked to relationship supporting functions, despite these having previously been linked to Non-Duchenne smiles [1], [31], pressed lips and dimpler [34], [35].

Second, we add verbal signals to the prediction framework and found that they reinforce a pattern of specificity rather than generality: dialogue acts that convey intentions to distance from the relationship to the interviewer and motivational language mark intrapersonal regulation, consistent with the approach-distance framework where a focus on own well-being is associated with abandoning a social relationship that elicits a self-concept threat. Interestingly, self-referential language was negatively associated with intrapersonal regulation, and rather marks outward relational functions (both relationship supporting and compromising/abandoning).

Third, we found that some smile functions vary by situational or temporal context: Intrapersonal and masking smiles were more likely in the first shame-eliciting situation, suggesting that applicants initially focus on regulating their own well-being, while in later situations they shift toward adjusting the relationship with the interviewer.

Fourth, and most critically, we quantify the between-person variability that prior work could not address: even after accounting for all observable signals, a substantial share of the differences in smile function is attributable to the individual displaying the smile. But not all functions were equally person-dependent: masking and relationship-adjusting functions vary more across individuals, while relationship-supporting functions show less between-person variation. This may explain why classification approaches, both in the original [1] and our multimodal approach, fail to consistently perform above baseline: because individuals differ systematically in how they resolve the approach-distance dilemma through smiling, the same observable features can serve different regulatory functions across persons, precluding universal decoding. These individual differences are consistent with the self-concept protection mechanisms described in shame research [5], [19]. When faced with the same evaluative threat, some individuals may defend their self-concept through masking smiles, others through contemptuous smiles, reflecting different shame regulation strategies [7] shaped by individual habits [4], [23]. This also explains why credible predictor effects coexist with machine learning classification failure: the regression model identifies credible verbal and nonverbal effects by accounting for person-level baselines, demonstrating that these features carry information, but the SVM (lacking person-specific baseline parameters) cannot account for between-person variability in the leave-one-subject-out evaluation. The classification failure is thus more likely attributable to between-person variability than to inadequate features.

Notably, adding observable predictors did not reduce but concentrated the between-person differences, indicating that predictors account primarily for variation across smile instances within the same person, leaving person-level baseline differences largely unchanged. Furthermore, individuals who smiled for self-regulatory reasons also tended to smile to mask emotions, suggesting a shared regulatory disposition at the person level. However, these variance decompositions establish *that* individuals differ, not *why* – they do not identify which person-level characteristics drive the differences.

Additionally, we found that neither machine learning classification nor human raters can consistently assign smile functions to observed smiles. The poor human annotation performance strengthens the assumption that the information required to distinguish smile functions is not reliably available in externally observable expressions, given that trained raters cannot agree with ground truth (functions that have been self-reported by the smiling individual), despite high mentalizing abilities, access to recordings of the shame-eliciting situations, and being familiarized with the full interaction context and smile functions. Even the most recognizable functions remained below the threshold for tentative agreement [70]. This reflects a genuine ambiguity that arises when the mapping from internal state to external expression varies across individuals [23], and when social display rules and emotion regulation strategies mask the connection between what is felt and what is shown [53]. The between-person variability we document may offer an explanation: different people may use the same observable cues to serve different regulatory functions, and without person-specific knowledge, neither computational nor human raters can resolve this ambiguity.

Our results also align with the approach-distance dilemma that characterizes shame-eliciting situations [21], [22] where individuals balance relationship maintenance and self-protection, with outcomes shaped by emotion regulation strategies [4], [23]. These findings suggest that the regulatory strategies individuals employ in such dilemmas reflect dispositions that produce systematic differences across individuals.

Overall, we contribute to the theoretical debate on emotion recognition and whether emotional expressions are universal [8]. The findings support a nuanced position: some observable signals show smile function-specific associations, yet between-person heterogeneity limits the generalizability of these associations across individuals. This also challenges current affective computing approaches [1], [56], [57] that focus on optimized features and model architecture. Future affective computing approaches should account for individual tendencies regarding, for instance, expressivity, emotion regulation strategies and social values [23], addressing the between-person differences documented in our model.

Limitations and Future Work. Although the human annotations suggest genuine ambiguity, different datasets and models may yield different estimates. The small sample and between-person variability limited the SVM's ability to generalize. While Bayesian regression models are suited for small samples, wide credible intervals reflect imprecise estimates that larger samples would narrow. Generalizability to other populations and socio-cultural contexts remains to be established. Notably, smile functions were derived from retrospectively self-reported internal experiences, which may be restricted due to emotion regulation processes [6], [73], altered through memory reconstruction [74], and are challenging for external annotators to judge. Future models may benefit from adding prosodic and phonological features [25], [75]. Also, how verbal and nonverbal signals unfold before, during, and after a smile deserves further investigation, as our exploratory analysis suggests timing-specific effects. Finally, between-person differences invite direct investigation: individual variables such as habitual emotion regulation strategies and social values [4], [23] may explain part of the observed variance.

A. Conclusion

This study provides evidence that internal smile functions cannot be reliably inferred from observable signals alone. Machine learning classifiers failed to beat trivial baselines, human raters reached only chance-level agreement, and regression models showed that outcomes were driven mainly by between-person differences in baseline smile tendencies. This between-person heterogeneity is not a limitation; it is a finding with direct implications for how emotion recognition systems should be designed. We propose that a reliable interpretation of smile functions from external observation requires accounting for who is smiling, not just what is visible.

Ethical Impact Statement

This study extends the corpus of smiles in shame-eliciting social situations originally established in [1] by adding comprehensive verbal annotations to existing audio-visual recordings. All underlying data collection procedures received prior approval from the Ethical Review Board of the authors' home institution, and participants provided informed consent for the capture and subsequent research use of their facial videos and voice recordings, as well as for the eventual publication of these materials in an anonymized form after removal of personally identifiable information. The corpus used in this study contains personally identifiable information; access to the original recordings was restricted to authorized research personnel bound by confidentiality agreements. The present work analyzes the original materials directly (e.g., for human annotation of smile functions and verbal coding). Public dissemination is restricted to a de-identified, processed version from which personally identifiable details have been removed.

To derive smile-related verbal annotations, we pre-coded dialogue acts using an open-source large language model. To protect participant privacy, all LLM inference ran in a protected enclave with end-to-end encrypted traffic, ensuring that no participant data were shared with third parties [67]. A human coder reviewed and corrected all automated outputs, making final coding decisions. Additionally, four trained raters independently annotated smile functions from the original video recordings. All individuals involved in data handling, coding, and annotation were bound by institutional confidentiality agreements.

The shame-elicitation paradigm required a simulated job interview with a virtual agent. To preserve ecological validity, participants were informed that they would interact with an autonomous, adaptive interviewer, when in fact an experimenter remotely controlled the agent's scripted behaviors while covertly monitoring the session. This deception was essential to elicit genuine emotional responses central to the research objectives. Following the interaction, a qualified psychologist conducted structured post-interviews regarding participants' subjective affective experiences and the regulatory functions of their smiles. Throughout the session, rapport-building techniques [59] were employed, regular welfare checks were performed, and a comprehensive debriefing disclosed the scripted nature of the interaction. Participants retained the right to withdraw at any time without penalty. Before dismissing them, the experimenter ensured that participants' well-being has not been impaired through the study.

The analyzed sample comprises 27 German psychology students (three exclusions from an initial set of 30 due to insufficient audio quality). While this population offers enhanced validity for introspective self-report data [76], its demographic and cultural homogeneity limits generalizability. Our findings further underscore substantial between-person variability in how internal experiences link to observable expressions, suggesting that universal inference models may be insufficient and that individualized approaches are needed for ecologically valid prediction.

Finally, this work probes the inferential gap between external multimodal signals and internal psychological states. Although our results demonstrate fundamental limits in decoding smile functions without intrapersonal information, the underlying technology raises significant privacy concerns. Automated inference of emotional states and social intentions from combined verbal and nonverbal signals could expose deeply private information, potentially compromising social standing, personal privacy, and psychological welfare. Transparent disclosure of system capabilities, data practices, and associated risks is therefore a prerequisite for any deployment; observation through such systems must never occur without the knowledge and explicit agreement of those observed.

References

[1] M. Hladký, R. Guerra, X. Cang, K. Maclean, P. Gebhard, and T. Schneeberger, "Modeling the 'kiss my ass'-smile: Appearance and functions of smiles in negative social situations," in *12th International Conference on Affective Computing and Intelligent Interaction (ACII). Glasgow, UK*, 2024.

[2] T. Schneeberger, M. Hladký, A.-K. Thurner, A. Heimerl, T. Baur, E. André, J. Volkert, and P. Gebhard, "Towards a deeper modeling of emotions: The DEEP method and its application on shame," in *Conference on Affective Computing and Intelligent Interaction*, 2021.

[3] T. Schneeberger, M. Scholtes, B. Hilpert, M. Langer, and P. Gebhard, "Can social agents elicit shame as humans do?" in *Conference on Affective Computing and Intelligent Interaction*, 2019, pp. 164–170.

[4] J. J. Gross and O. P. John, "Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being." *Journal of Personality and Social Psychology*, vol. 85, no. 2, p. 348–362, 2003.

[5] M. Lewis, "Self-conscious emotions: Embarrassment, pride, shame, and guilt." in *Handbook of Emotions*, M. Lewis, J. M. Haviland-Jones, and L. Feldman Barrett, Eds. Guilford, 2008, vol. 3, p. 742–756.

[6] J. J. Gross, *Handbook of emotion regulation*. Guilford, 2024.

[7] D. L. Nathanson, *Shame and pride*. Norton, 1994.

[8] L. Feldman Barrett, *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[9] U. Moser and I. Von Zeppelin, "Die Entwicklung des Affektsystems," *Psyche*, vol. 50, no. 1, pp. 32–84, 1996.

[10] L. F. Barrett, "Are emotions natural kinds?" *Perspectives on psychological science*, vol. 1, no. 1, pp. 28–58, 2006.

[11] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: A meta-analytic review," *Behavioral and Brain Sciences*, vol. 35, no. 3, pp. 121–143, 2012.

[12] J. J. Gross, O. P. John, and J. M. Richards, "The dissociation of emotion expression from emotion experience," *Personality and Social Psychology Bulletin*, vol. 26, no. 6, pp. 712–726, 2000.

[13] K. Niven, "Interpersonal emotion regulation," in *Handbook of Emotion Regulation*. Guilford, 2024, pp. 153–166.

[14] M. Hladký, T. Schneeberger, and P. Gebhard, "Understanding shame signals: Functions of smile and laughter in the context of shame," in *9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2021.

[15] S. S. Dickerson, T. L. Gruenewald, and M. E. Kemeny, "When the social self is threatened: Shame, physiology, and health," *Journal of Personality*, vol. 72, no. 6, pp. 1191–1216, 2004.

[16] R. F. Baumeister, Ed., *The self in social psychology*, ser. The self in social psychology. Psychology Press, 1999.

[17] C. E. Izard, *Human emotions. Emotions, personality, and psychotherapy*. Plenum, 1977.

[18] W. K. Hahn, "The experience of shame in psychotherapy supervision." *Psychotherapy*, vol. 38, no. 3, p. 272, 2001.

[19] D. L. Nathanson, "Shame and the affect theory of Silvan Tomkins." in *The widening scope of shame*, A. P. Morrison and M. R. Lansky, Eds. The Analytic Press, 1997.

[20] J. P. Tangney, S. A. Burggraf, and P. E. Wagner, "Shame-proneness, guilt-proneness, and psychological symptoms." in *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride*. Guilford, 1995.

[21] K. Lewin, *A Dynamic Theory of Personality*. McGraw-Hill, 1935.

[22] A. J. Elliot, "The Hierarchical Model of Approach-Avoidance Motivation," *Motivation and Emotion*, vol. 30, no. 2, pp. 111–116, 2006.

[23] M. Hladký, B. Han, J. Gratch, P. Gebhard, and T. Schneeberger, "Steal and Smile: Impact of Emotion Regulation and Social Values on Smiles in Social Conflicts," in *13th International Conference on Affective Computing and Intelligent Interaction (ACII). Canberra, AU*, 2025.

[24] S. Paulmann and M. D. Pell, "Is there an advantage for recognizing multi-modal emotional stimuli?" *Motivation and Emotion*, vol. 35, no. 2, pp. 192–201, 2011.

[25] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS ONE*, vol. 6, no. 11, p. e27256, 2011.

[26] M. Zellers, J. Gorisch, and D. House, "Temporal relationships between speech and hand gestures in the vicinity of potential turn boundaries in german and swedish conversation," *Language and Cognition*, 2025.

[27] B. L. Fredrickson and R. W. Levenson, "Positive emotions speed recovery from the cardiovascular sequelae of negative emotions," *Cognition & emotion*, vol. 12, no. 2, pp. 191–220, 1998.

[28] C. C. Neuhoff and C. Schaefer, "Effects of laughing, smiling, and howling on mood," *Psychological reports*, vol. 91, pp. 1079–1080, 2002.

[29] E. Krumhuber and A. Kappas, "More What Duchenne Smiles Do, Less What They Express," *Perspectives on Psychological Science*, vol. 17, pp. 1566–1575, 2022.

[30] U. Hess, M. G. Beaupré, N. Cheung *et al.*, "Who to whom and why–cultural differences and similarities in the function of smiles," *An empirical reflection on the smile*, vol. 4, p. 187, 2002.

[31] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of nonverbal behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[32] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed," *Nonverbal behavior*, vol. 33, no. 1, pp. 17–34, 2009.

[33] W. Ruch, "Extraversion, alcohol, and enjoyment," in *What the face reveals*, P. Ekman and E. L. Rosenberg, Eds. Oxford, 1997.

[34] M. Rychlowska, R. E. Jack, O. G. B. Garrod, P. G. Schyns, J. D. Martin, and P. M. Niedenthal, "Functional Smiles: Tools for Love, Sympathy, and War," *Psychological Science*, vol. 28, no. 9, pp. 1259–1270, 2017.

[35] J. D. Martin, A. Wood, W. T. L. Cox, S. Sievert, R. Nowak, E. Gilboa-Schechtman, F. Zhao, Z. Witkower, A. T. Langbehn, and P. M. Niedenthal, "Evidence for Distinct Facial Signals of Reward, Affiliation, and Dominance from Both Perception and Production Tasks," *Affective Science*, vol. 2, no. 1, pp. 14–30, 2021.

[36] M. Burdett, M. Pino, N. Moghaddam, and T. Schröder, ""It sounds silly now, but it was important then": Supporting the significance of a personal experience in psychotherapy," *Journal of Pragmatics*, vol. 148, pp. 12–25, 2019.

[37] J. B. Torre and M. D. Lieberman, "Putting feelings into words: Affect labeling as implicit emotion regulation," *Emotion Review*, vol. 10, no. 2, pp. 116–124, 2018.

[38] S. J. Torrisi, M. D. Lieberman, S. Y. Bookheimer, and L. L. Altshuler, "Advancing understanding of affect labeling with dynamic causal modeling," *NeuroImage*, vol. 82, pp. 481–488, 2013.

[39] P. Ferré, I. Fraga, and J. A. Hinojosa, "The interplay between language and emotion: A narrative review," *Cognition and Emotion*, vol. 39, no. 7, pp. 1418–1445, 2025.

[40] T. Meier, R. L. Boyd, J. W. Pennebaker, M. R. Mehl, M. Martin, M. Wolf, and A. B. Horn, ""LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE- LIWC2015," 2018.

[41] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[42] W. Weintraub, *Verbal behavior: Adaptation and psychopathology*. Springer, 1981.

[43] R. A. Simmons, D. L. Chambless, and P. C. Gordon, "How do hostile and emotionally overinvolved relatives view relationships? What relatives' pronoun use tells us," *Family Process*, vol. 47, no. 4, pp. 405–419, 2008.

[44] R. B. Slatcher, S. Vazire, and J. W. Pennebaker, "Am "I" more important than "we"? Couples' word use in instant messages," *Personal Relationships*, vol. 15, no. 4, pp. 407–424, 2008.

[45] J. W. Pennebaker, T. J. Mayne, and M. E. Francis, "Linguistic predictors of adaptive bereavement," *Journal of Personality and Social Psychology*, vol. 72, no. 4, pp. 863–871, 1997.

[46] J. R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press, 1979.

[47] H. Bunt, "Dialogue pragmatics and context specification," *Abduction, Belief and Context in Dialogue*, pp. 81–119, 2000.

[48] D. R. Traum and S. Larsson, "The Information State Approach to Dialogue Management," in *Current and New Directions in Discourse and Dialogue*, J. van Kuppevelt and R. W. Smith, Eds. Dordrecht: Springer Netherlands, 2003, pp. 325–353.

[49] M. Yeomans, *A Practical Guide to Conversation Research: How to Study What People Say to Each Other*. Cambridge University Press, 2023.

[50] R. Clift, "Don't make me laugh: Responsive laughter in (dis)affiliation," *Journal of Pragmatics*, vol. 100, pp. 73–88, 2016.

[51] J. Heritage and S. Clayman, *Conversation Analysis and Institutional Talk*. Cambridge University Press, 2005.

[52] M. Mongrain and L. C. Vettese, "Conflict over Emotional Expression: Implications for Interpersonal Communication," *Personality and Social Psychology Bulletin*, vol. 29, no. 4, pp. 545–555, 2003.

[53] A. R. Hochschild, *The Managed Heart: Commercialization of Human Feeling*. University of California Press, 2015.

[54] E. Goffman, *The Presentation of Self in Everyday Life*. Anchor Books, 2023.

[55] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[56] B. Han, S. Lei, C. Yau, and J. Gratch, "Knowledge-based emotion recognition using large language models," in *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2024.

[57] R. Kosti, J. M. Álvarez, A. Recasens, and À. Lapedriza, "Context based emotion recognition using EMOTIC dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1685–1700, 2020.

[58] P. Gebhard, T. Schneeberger, T. Baur, and E. André, "Marssi: Model of appraisal, regulation, and social signal interpretation," in *Int. Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 497–506.

[59] T. Schneeberger, M. Hladký, A.-K. Thurner, J. Volkert, A. Heimerl, T. Baur, E. André, and P. Gebhard, "The deep method: Towards computational modeling of the social emotion shame driven by theory, introspection, and social signals," *IEEE Transactions on Affective Computing*, pp. 1–16, 2023.

[60] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum, "Towards an ISO standard for dialogue act annotation," *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[61] P. Mayring, *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Beltz, 2014.

[62] P. Törnberg, "Best Practices for Text Annotation with Large Language Models," *Sociologica*, vol. 18, no. 2, pp. 67–85, 2024.

[63] M. T. Whitty, "How to Conduct AI-Assisted (Large Language Model-Assisted) Content Analysis in Information Science and Cyber Security Research," *Electronics*, vol. 14, no. 20, p. 4104, 2025.

[64] M. Vikan, R. Aryan, M. S. Kannelønning, M. A. Riegler, and S. O. Danielsen, "Reflecting on LLM Support in Reflexive Thematic Analysis: An Exploratory Study," *Qualitative Health Research*, vol. 36, no. 2-3, pp. 191–205, 2025.

[65] X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao, "Human-llm collaborative annotation through effective verification of llm labels," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024.

[66] M. Farjam, H. Meyer, and M. Lohkamp, "A Practical Guide and Case Study on How to Instruct LLMs for Automated Coding During Content Analysis," *Social Science Computer Review*, p. 08944393251349541, 2025.

[67] Venice AI, Inc., "Venice.ai: Private and uncensored ai platform," Software, 2026, accessed: 2026-04-22. [Online]. Available: https://venice.ai

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*,

"Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[69] L. A. Kasper, S. Hauschild, A. Berning, J. Holl, and S. Taubner, "Development and validation of the Mentalizing Emotions Questionnaire: A self-report measure for mentalizing emotions of the self and other," *PLOS ONE*, vol. 19, no. 5, p. e0300984, 2024.

[70] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 3rd ed. Sage Publications, 2011.

[71] G. Marzi, M. Balzano, and D. Marchiori, "K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient," *MethodsX*, vol. 12, p. 102545, 2024.

[72] J. J. Hox, *Multilevel Analysis: Techniques and Applications*, 2nd ed. Routledge, 2010.

[73] L. Feldman Barrett, "Feelings or words? understanding the content in self-report ratings of experienced emotion," *Journal of Personality and Social Psychology*, vol. 87, no. 2, pp. 266–281, 2004.

[74] H. Roediger and K. DeSoto, "The psychology of reconstructive memory." pp. 50–55, Dec. 2015.

[75] A. Majid, "Current emotion research in the language sciences," *Emotion Review*, vol. 4, no. 4, pp. 432–443, 2012.

[76] J. A. Hall, "Psychological-mindedness: A conceptual model," *American Journal of Psychotherapy*, vol. 46, no. 1, pp. 131–140, 1992.