

Not All Subjectivity Is the Same!

Defining Desiderata for the Evaluation of Subjectivity in NLP

Urja Khurana[🔥], Michiel van der Meer^{🔥*}, Enrico Liscio^{🔥*},

Antske Fokkens[🔥], Pradeep K. Murukannaiah[🔥]

[🔥]Department of Intelligent Systems, Delft University of Technology

[🔥]Leiden Institute of Advanced Computer Science (LIACS), Leiden University

[🔥]Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam

{u.khurana, e.liscio, p.k.murukannaiah}@tudelft.nl,

m.t.van.der.meer@liacs.leidenuniv.nl, antske.fokkens@vu.nl

Abstract

Subjective judgments are part of several NLP datasets and recent work is increasingly prioritizing models whose outputs reflect a diversity of perspectives. Such responses allow us to shed light on minority voices, which are frequently marginalized or obscured by dominant perspectives. It remains a question whether our evaluation practices align with these models' objectives. This position paper proposes *seven* evaluation desiderata for subjectivity-sensitive models, rooted in how subjectivity is represented in NLP data and models. The desiderata are constructed in a top-down approach, keeping in mind the user-centric impact of such models. We scan the experimental setup of 64 papers and show that various aspects of subjectivity are still understudied: the distinction between ambiguous and polyphonic input, whether subjectivity is *effectively* expressed to the user, and a lack of interplay between different desiderata, amongst other gaps.

1 Introduction

For several NLP tasks, such as answering general knowledge questions, there is a *factual* or *correct* response that is generally accepted. However, for some tasks, the “truth” depends on the context or individual interpretation. Consider, for example, the task of determining whether an online comment is offensive. There is no universally agreed notion of *offensiveness* or *hate speech* (Vidgen et al., 2019; Fortuna et al., 2020; Leonardelli et al., 2021; Khurana et al., 2022). Instead, the sociocultural background and experiences of the readers influence their judgments (Sap et al., 2022; Goyal et al., 2022). Similarly, judgments depend on a personal interpretation in many other tasks such as natural language inference (Pavlick and Kwiatkowski, 2019; Nie et al., 2020), semantic annotation (Sommerauer et al., 2020), and understanding indirect

answers (Damgaard et al., 2021). We refer to such tasks that have multiple plausible labels rather than a single valid ground-truth as *subjective*.

As language models are integrated into societal applications, they increasingly engage with subjective concepts such as values, norms, and arguments (Forbes et al., 2020; Lourie et al., 2021; Liscio et al., 2023; Van Der Meer et al., 2024b). Therefore, their behavior must align with varying human preferences (Sorensen et al., 2024) and reflect multiple viewpoints (Cabitza et al., 2023; Plank, 2022). Yet, such models have been shown to collapse to a specific view (Hayati et al., 2024), typically as a consequence of how labels are presented (aggregated to a single majority label) or due to prevalent preference-tuning paradigms (such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023)) that optimize for the majority vote or general preference (Casper et al., 2023). Obfuscating minority or diverging viewpoints is problematic when deploying models in real-world settings, running the risk of further disenfranchising marginalized communities (Xu et al., 2021).

Subjectivity-sensitive models reflect variation in judgment and enable the representation of minority perspectives. As the NLP field pushes for such models (Ovesdotter Alm, 2011; Basile et al., 2021; Plank, 2022; Cabitza et al., 2023; Wang et al., 2025b), a critical look at their evaluation practices is necessary. Several aspects affect the evaluation of subjectivity-sensitive models, including the ideals we outline, the extent to which subjective evaluation metrics capture overall trends or fine-grained model behavior (Chen et al., 2024a; Rizzi et al., 2024), and the type of task (classification vs. generation). To this end, we address the following question: **which aspects of subjectivity should evaluation metrics capture?**

To address this question, in this position paper, we first review how subjectivity is represented in the data and incorporated into NLP models, high-

*Equal contribution.

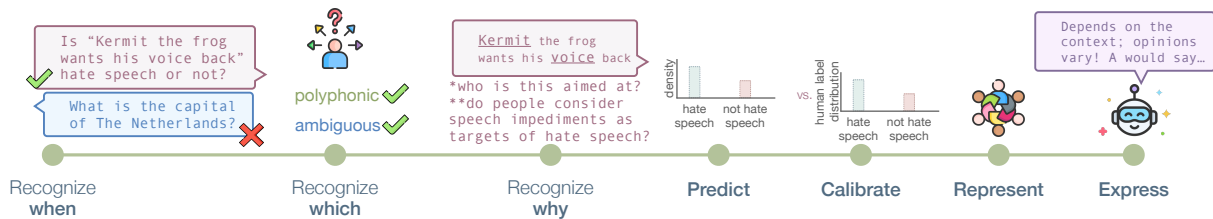


Figure 1: The sequential nature of our proposed desiderata for evaluating a subjectivity-sensitive model.

lighting a previously underexplored aspect: the distinction between ambiguity (input that can warrant multiple interpretations) and polyphony (input that has multiple valid perspectives). These insights further drive our top-down approach, in which we propose a set of sequential desiderata to evaluate subjectivity-sensitive models (outlined in Figure 1). We identify aspects not yet addressed in the literature by drawing on insights from related fields. Finally, we map these desiderata onto current evaluation practices, revealing gaps and opportunities for future work.

Our key contributions are threefold: (1) desiderata for evaluating subjectivity-sensitive models, (2) a systematic and indicative categorization of current evaluation metrics, and (3) identified gaps and opportunities.

2 Subjectivity in Data

To establish a fitting evaluation paradigm, we start by describing how subjectivity and related concepts are interpreted in the NLP field. Next, we discuss what the *sources* of subjectivity are. Finally, we reflect on the types of information to incorporate in a subjectivity-sensitive dataset.

2.1 Subjectivity and Related Concepts

Subjectivity pertains to the existence of multiple valid ground-truths (Vardomskaya, 2018; Rottger et al., 2022; Plank, 2022). NLP literature also discusses other concepts related to subjectivity. *Perspectivism* (Cabitza et al., 2023; Frenda et al., 2025) recognizes that ground-truth depends on individual perspectives. This has taken shape as a *perspectivist* framework for representing different perspectives in data: *weak* when multiple perspectives are considered only during annotation, and *strong* when extending this to incorporating multiple perspectives in the model output as well. Subjective interpretations can lead to *disagreement* (Uma et al., 2021b) when different individuals have opposing views on a specific input. Plank (2022)

uses the term *human label variation* to refer to instances with different plausible perspectives, since *disagreement* could imply that at least one judgment is incorrect. *Disagreement* can also arise in factual matters, capturing annotator *errors* due to misunderstanding (misinformation or misinterpreting annotation guidelines) or attention drifts.

2.2 Sources of Subjectivity

Where do multiple subjective viewpoints or beliefs stem from? Aroyo and Welty (2015) suggest three potential sources of disagreement in their Triangle of Reference framework that maps out the process of interpreting an input: *task ambiguity*, *input ambiguity*, and *annotator attitude* (Jiang and de Marnaffe, 2022; Liu et al., 2023). For *task ambiguity*, imprecise task definitions and annotation guidelines may (un)intentionally leave enough space for an annotator to fill the gaps through their own interpretation (Rottger et al., 2022). For instance, in “*Berlin decided ...*”, the entity *Berlin* can be annotated as a location or an organization. Without explicitly guiding annotators in such scenarios (in this case, where a city name is used to refer to a government), unintended disagreement can arise. Even when the task is clearly defined, *input ambiguity* may still arise when more context is needed for an accurate judgment. Consider an input that contains a reclaimed slur; whether it is hate speech requires the knowledge of who is behind the piece of text. Finally, *annotator-specific factors* such as an annotator’s demographic, emotional state, and experience can yield different valid judgments (Sap et al., 2022), particularly when there is no universally agreed-upon ground-truth.

2.3 Incorporating Subjectivity in Data

To train and evaluate models that can handle subjectivity, adequate datasets should be available. Conventionally, NLP datasets are built by collecting n annotations per sample and aggregating them into a single “gold” label, e.g., through majority voting. However, when there is variation in anno-

tator response, this results in a loss of information on human disagreement. This has led to calls for the release of raw annotations (Prabhakaran et al., 2021; Cabitza et al., 2023; Pandita et al., 2026), with recent work showing the importance of having a large number of annotations per sample (Khurana et al., 2024; Gruber et al., 2024).

To model the opinion of an individual (annotator), we would need demographic information (Bender and Friedman, 2018; Prabhakaran et al., 2021) and a sufficient amount of annotated samples. Disagreement can also stem from annotation mistakes (e.g., due to attention drift), requiring methods (Gordon et al., 2021; Weber-Genzel et al., 2024; Jinadu and Ding, 2024; Ivey et al., 2025) and metrics (Dumitrache et al., 2018; Abercrombie et al., 2025) that can distinguish such *noise* from actual signal (i.e. ways to distinguish variation due to errors from variation reflecting differences in perspective) and indicate annotator reliability.

The part of the subjective spectrum that is captured is influenced by the pool of annotators. Smarter annotation techniques better capture how an input is perceived, e.g., by having a variety of demographics annotate a specific sample (Pei and Jurgens, 2023; Van Der Meer et al., 2024a), selecting annotators of a specific target group (Goyal et al., 2022), or a framework for subjectivity throughout the annotation stages (Díaz et al., 2022).

3 Dimensions of Subjective Input

What do the sources of subjectivity mean for a subjectivity-sensitive model? We take inspiration from the Triangle of Reference framework (Aroyo and Welty, 2015) discussed earlier. Where *task ambiguity* plays an overarching role when collecting ground-truth labels across multiple samples, *input ambiguity* and *annotator attitude* can primarily occur at a sample level. These two sources require the model to handle subjectivity differently, an aspect that has not yet been addressed in previous research. We characterize model behavior when dealing with different sources of subjectivity on two dimensions: **ambiguity** and **polyphony** (Figure 2).

When an input is non-ambiguous and monophonic, a model can directly incorporate the objective ground-truth. However, modeling scenarios involving subjective factors is more challenging as the following subsections illustrate.

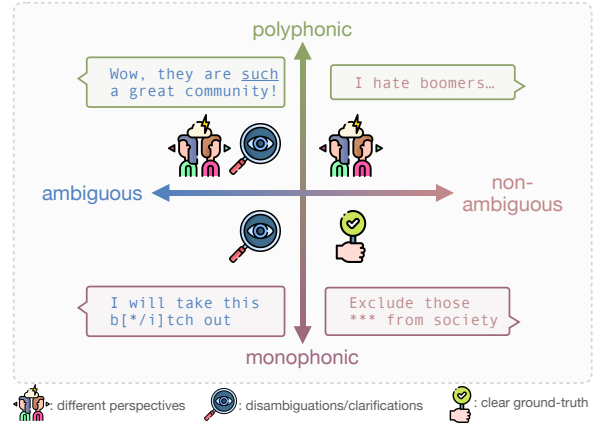


Figure 2: Dimensions of sources of subjectivity at a **sample-level**: *ambiguity* (x-axis) and *polyphony* (y-axis). Each quadrant shows the desired model output (as icons) with examples situated in hate speech detection.

3.1 Ambiguous ↔ Non-Ambiguous

An input is *ambiguous* when it can warrant different interpretations because the information (e.g., context) needed to make a judgment is incomplete. Inputs that can be interpreted without filling information gaps are *non-ambiguous*.

Ambiguous. When faced with an ambiguous input, a model should identify the spectrum of different possible interpretations and disambiguations. For instance, in “I will take this b[*i]tch out” the ambiguity lies in who the *b-word* refers to; the slur or the dog.

Non-ambiguous. A non-ambiguous input removes the requirement for a model to understand the different interpretations. In such a situation, the model can directly give a suitable response based on the exact information outlined in the input.

3.2 Polyphonic ↔ Monophonic

Independent of ambiguity, certain inputs may not have an agreed-upon ground truth, allowing for judgments based on an individual’s interpretation. Personal taste, demographics, background, experience, and other factors influence this interpretation. We consider such cases to be *polyphonic*: different perspectives co-existing. Where there is unanimity, we consider that to be *monophonic*.

Polyphonic. A polyphonic input is open to multiple valid perspectives, which the model should capture. Take for example the statement “I hate boomers”; the meaning is clear but the polyphony lies in whether boomers are considered targets of

hate speech. A model that can capture polyphony for the right reasons will be able to generalize well in the real world. Thus, the model should recognize *what* aspect(s) of the input enable different perspectives and *why*, e.g., is it the gender or culture that potentially influences the opinion or are there other factors? Not every different perspective has to be attributed to specific demographics, as deviations can also be due to individual differences.

Monophonic. When an input has a single valid belief, often connected to factuality, the model can directly predict the ground-truth without the need to consider a variety of potential perspectives. For instance, the phrase “*Exclude those *** from society*” should not warrant any subjective interpretation due to its overt form of hate speech.

3.3 Ambiguous and Polyphonic

Ambiguity and polyphony can be intertwined. Consider an input that lacks context *and* warrants multiple perspectives, e.g., “*Wow, they are such a great community*”. Here, the ambiguity lies in the sentiment of the sentence and the polyphony in the gravity of the (potential) negative sentiment. In such a scenario, a model is expected to capture the appropriate disambiguation and clarification questions, and the different possible perspectives.

4 Subjectivity in Models

How a model expresses subjectivity influences its evaluation. Thus, we briefly discuss ways of expressing and instilling subjectivity in models. Most papers on subjectivity focus on modeling, which we refer to for detailed discussions (Frenda et al., 2025; Xu and Jurgens, 2026). We notice a clear lack of work surrounding how to evaluate subjectivity, which we expound on in Sections 5 and 6.

4.1 Expressing Subjectivity

To move away from rigid responses in line with the majority preference, the model’s output should reflect the spectrum of different perspectives or be aligned to an individual’s perspective; *pluralistically aligned*. Sorensen et al. (2024) operationalize this as follows: (1) **Distributional**: the model outputs the distribution of perspectives. This can be expressed in the logits or softmax distribution of the model, or directly in the generated response (Meister et al., 2025). (2) **Overton**: the output summarizes different *reasonable* perspectives. The summary should accurately reflect the window of

perspectives. (3) **Steering**: the model considers multiple perspectives and values and steers its response to the user’s. This is closely related to *personalization*, where preferences go beyond values and include matters such as *textual* or *linguistic* style (Zhang et al., 2025b).

4.2 Instilling Subjectivity

We categorize approaches based on whether subjectivity is instilled when training or fine-tuning, during specific alignment steps, or during inference. At **inference-time**, integrating subjectivity is usually more computationally efficient than at the training/fine-tuning stage, particularly in the case of LLMs. A popular approach is persona prompting (Santurkar et al., 2023; Deshpande et al., 2023), where the model is instructed in a prompt to impersonate a specific *persona*. This can span different demographics where the persona could steer toward a specific perspective or provide a spectrum of multiple perspectives. Typically, for subjective modeling during **fine-tuning** we use *soft* labels, i.e., the distribution of class annotations for a single sample (Jamison and Gurevych, 2015; Uma et al., 2021a). Instead of optimizing for a single (aggregated) *hard* label, we minimize the distance between the estimated and the actual distribution. For this we use distance-based loss functions such as *cross-entropy*, *Kullback-Leibler divergence*, or *Jensen-Shannon divergence*. We can also learn from individual annotators’ labels to make sets of predictions that constitute annotation distributions (Rodrigues and Pereira, 2018). **Preference tuning** methods for large language models primarily model majority preference (Casper et al., 2023) but strategies are being developed to reflect diverse preferences (Zhao et al., 2024; Poddar et al., 2024).

5 Desiderata for Subjective Models

After outlining the desired model output when facing a subjective input, the natural next step is to evaluate its performance. *But what does it mean for a model to perform well on subjective input?*

To answer this question, we propose desiderata for a subjectivity-sensitive model. These desiderata are motivated by three overarching ideals rooted in our preceding discussion of how subjectivity is represented in NLP data and models: (1) representation of all, including minority voices, (2) based on the correct sources of subjectivity in the input, and (3) the representation of subjectivity (in data

and models) generalizing to the real world.

D1. Recognize when subjective modeling is necessary. As we highlight in Section 2.1, inputs may elicit multiple valid perspectives. For a model to make predictions that reflect this diversity, the first step is to **recognize when** an input needs to be treated subjectively versus objectively. Even when a task is generally considered to be subjective, this step is vital—there will be cases where raters will have consensus, e.g., when dealing with a severe case of hate speech; predicting an aggregate estimate would be sufficient in that case.

D2. Recognize which type of subjective modeling is needed. As we describe in Section 3, model behavior depends on the type of subjective input. Thus, a model should **recognize which** type of subjectivity the input contains: ambiguity, polyphony, or both.

D3. Recognize why there is subjectivity in the input. Building on the desired model behavior we outline in Section 3, a model should **recognize why** there is ambiguous or polyphonic subjectivity in the input. Using this information, the model can delineate the different disambiguations, clarifications, or perspectives to resolve. This facilitates better predictions and more reliable generalization to unseen scenarios.

D4. Predict subjectivity accurately. After recognizing the source of the subjectivity, a model should predict the subjectivity correctly. A model can **predict** subjectivity in different ways, as we discussed in Section 4.1.

D5. Calibrate the output with respect to the frequency of opinions. As illustrated in Section 4.2, certain models are trained to estimate the distribution of human judgment for an input. Depending on how subjectivity is expressed (distributional or overton), the predicted distribution should be **calibrated** to the actual human judgment distribution. This applies both when a probability distribution is obtained directly from the model and when an overton response (where linguistic quantifiers should be well-calibrated) is generated.

D6. Represent all perspectives. The model should account for the minority voices that are usually lost in aggregation, by **representing** diverse viewpoints. Relevant perspectives should be decoupled from dominant perspectives. For example,

it should not only echo a Western perspective but also other cultures when the input warrants it.

D7. Express subjectivity without hampering user experience. A subjectivity-sensitive model carries with it the responsibility of *informing* a user of different perspectives. This mirrors the responsibility of showing users explanations of decision-making in the field of explainable AI (XAI). What counts as a “good” explanation depends on the user (Ehsan and Riedl, 2020). In a similar vein, we argue that subjectivity should be **expressed** in a way that does not hamper user experience. A response that enumerates all potential perspectives can overload the user.¹ Though overload can sometimes be prevented by structuring responses, this desideratum can also be at odds with **D6**. Ideally, the perspectives included in a response should be *useful*. However, this is challenging as what is useful is user- and context-dependent (e.g., to what extent do we consider minority opinions, even when *one* person holds it? Does a user find it important to expand their worldview? Or would they rather know more about perspectives from similar cultures? And is what a user wants always that what the user needs?). To summarize, usefulness is a complex notion and what this means in practice requires more research.

Practical considerations. Our desiderata can be evaluated sequentially and sometimes also independently. They allow for a mix-and-match approach depending on the nature and setup of the task at hand. Certain desiderata are connected by looking at the model output from different angles. We illustrate this complementary nature of desiderata **D4**, **D5**, and **D6** in Figure 3. Note that the model does not have to be explicitly designed to address all of the desiderata, but can capture them implicitly as well (e.g., a model does not have to be explicitly designed to output binary labels for **D1**, it can cover this desideratum implicitly when directly predicting the human judgment distribution). The application also influences to what extent all the desiderata should be covered by the model, as well as the order of certain desiderata. For instance, for an NLI model doing requirement analysis, **D6. Represent all perspectives** is not a priority. There could also be cases where the cause of subjectivity can be identified without knowing what exact type of subjectivity is being dealt with.

¹This excludes a *distributional* output from the model.

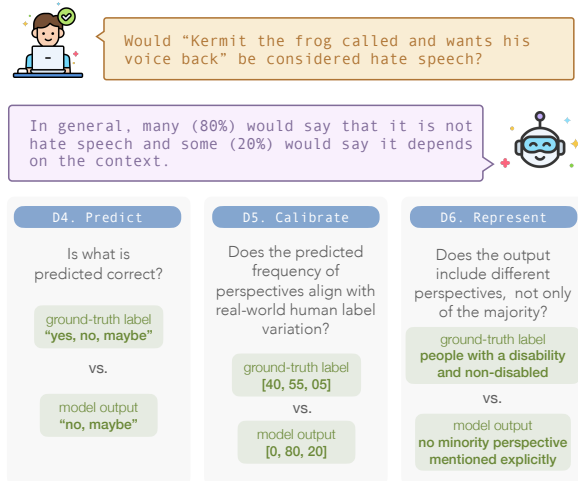


Figure 3: Complementary nature of the desiderata that specifically evaluate the model output from different angles: **D4. Predict**, **D5. Calibrate**, and **D6. Represent**. The example (including all the values for the labels) is for illustrative purposes only, showcasing what each desideratum captures. In this case the input text (taken from Schmidt and Wiegand (2017)) is highly contextual and could potentially be harmful to people with a speech disability. Each desideratum focuses on a different angle, though **D4** and **D5** can have some overlap if distance metrics are used for **D4**.

6 Subjectivity Evaluation Practices

In Section 5, we formulated the desiderata for a subjectivity-sensitive model in a top-down fashion. We now look bottom-up, exploring how these desiderata translate to concrete evaluation practices. We discuss existing metrics and analyze how current evaluation practices relate to our desiderata in both classification and free-text generation settings.

6.1 Existing Metrics

We first discuss metrics used to measure aspects of how well a model captures subjectivity in its predictions. We review related literature and identify the main metrics and evaluation practices.²

Predicting the full human judgment distribution.

Determining whether subjectivity is accurately predicted depends on how the model *expresses* it. When *disagreement* is represented as *soft* labels, we measure the distance between the soft labels and the softmax distribution of the model. Widely used metrics are (variations of) *KL-divergence*, *Jensen-Shannon Divergence*, *Mean Absolute Er-*

²Our aim is not to provide an exhaustive list of existing metrics, but to provide a snapshot of important metrics and evaluation methods to identify gaps the field can focus on.

ror, *Wasserstein*, or *Total Variation Distance* (Rizzi et al., 2024). To assess the spread of a single distribution, entropy $\mathcal{H}$ is used (Alies et al., 2025). Whether the model captures human disagreement, we can measure the correlation between the entropy of the soft label and the softmax distribution ($\mathcal{H}_{corr}$) (Uma et al., 2020). These metrics work for classification, but are not applicable when evaluating free-text generated responses. For generation, NLI models are used to determine whether the generated response covers the different necessary perspectives (Feng et al., 2024; Shetty et al., 2025).

Single label scenarios. Although not suitable to evaluate *soft* labels, *ACCURACY* and *F₁-SCORE* are still used in subjective scenarios, particularly when modeling an individual user (steering) or evaluating an aggregation of the subjective predictions to a majority label (Mostafazadeh Davani et al., 2022). Additionally, looking at specific minorities may reveal performance insights on the least represented annotators (Van Der Meer et al., 2024a). *F₁-SCORE* can also be used to determine whether a model is good at recognizing subjective input (Homayounirad et al., 2025), or turned into soft versions (Kurniawan et al., 2026).

Demographical bias. Certain metrics measure the extent to which a model represents (or is biased toward) specific demographic groups. This can be done by measuring the similarity in responses between a model and a specific country (Durmus et al., 2024) or demographic (Santurkar et al., 2023), or using a linear mixed-effects model (Sun et al., 2025). Jiang et al. (2025) propose *VALUE INEQUITY INDEX* for whether model responses are sensitive to an individual’s demographic or not.

Reliability. Reliability metrics such as *CORRELATION* and *inter-annotator agreement* (IAA) are used for a variety of analyses. IAA is used mainly to measure the agreement between the predictions of the model and the human prediction (Masud et al., 2024; Hoeken et al., 2025). Apart from $\mathcal{H}_{corr}$, most commonly, such metrics can be used to measure whether the behavior of an LLM globally corresponds to human ratings of items in the dataset (Sap et al., 2022; Ryan et al., 2024; Parappan and Henao, 2025; Chen et al., 2024a), to specific demographics (Giorgi et al., 2024; Pihulski and Kocoń, 2025) or for other fine-grained analyses (Guo et al., 2025; Luz de Araujo et al., 2025; Plepi et al., 2022; Giorgi et al., 2024; Székely et al., 2007).

Calibration. A model is deemed well-calibrated when its confidence scores match its empirical performance. For subjective tasks, this means predicted probabilities match the human judgment distribution. While DISTANCE METRICS already capture this, Baan et al. (2022) propose measuring on an instance-level the difference in entropy between the soft label and the model’s predicted distribution and whether the class rankings match.

Other. The following metrics do not fall into the categories above or are too infrequent to be a separate category. Santurkar et al. (2023) propose measuring *consistency*; does the model’s default output consistently reflect the same demographic across different topics? VISUALIZING important dynamics has also gained traction (Wright et al., 2024; Chen et al., 2024a). Alternatively, LLM judges are used to evaluate whether a steered response is better than no steering, represented through the WIN RATE: the proportion of samples for which the steered model is preferred (Chen et al., 2025; Tan and Lee, 2025; Jang et al., 2024). Finally, Samuel et al. (2025) propose PersonaScore, a human-aligned score that evaluates the generated free-text response according to persona agent-specific tasks.

6.2 Metrics and Desiderata in Papers

We analyze how the metrics are used in practice in relation to our desiderata. Based on the metrics we discussed, we distinguish the following categorization of metrics that are regularly used for evaluating aspects of subjectivity: DISTANCE METRIC, F_1 , ACCURACY, RELIABILITY, and OTHER.

We describe our literature search and annotation method in Appendix A. We found 64 papers and annotated them, indicating the category of metrics used, the evaluation objective, and the covered desiderata. Figure 4 shows our mapping. These visualizations are meant to give a general idea of evaluation practices; the distributions presented in the figures are based on an indicative (not exhaustive) literature search.

D1.Recognize when there is subjectivity is captured as a binary classification task (subjective or not) or as a continuous task (how subjective). In the binary case, F_1 (Wan et al., 2023; Homayounirad et al., 2025) and ACCURACY are used (Romberg, 2022). Depending on how the continuous label is represented, evaluation ranges from using a distance metric (Wan et al., 2023) to $\mathcal{H}_{corr}$.

D2.Recognize which type of subjectivity is not explicitly evaluated in the papers we considered. Few studies compare performance by subjectivity type by either manually labeling each sample’s subjectivity type (Jiang and de Marneffe, 2022; Sandri et al., 2023) or using datasets tailored to specific types of subjectivity (Feng et al., 2024).

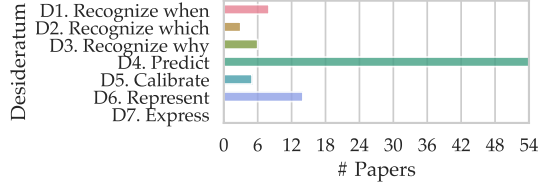
D3.Recognize why the input is subjective is evaluated in a single study, which tests whether a model can generate and recognize disambiguations in NLI (Liu et al., 2023). For generation, EDIT- F_1 (an F_1 -variant accounting for added and deleted unigrams between reference and generation) was used and ACCURACY for the recognition of disambiguations. Other papers conducted a qualitative analysis of keywords responsible for prediction using feature attribution (Wan et al., 2023; Muscato et al., 2025) or manually annotated the reasons for subjectivity and compared model performance accordingly (Sandri et al., 2023; Hoeken et al., 2025).

D4.Predict subjectivity accurately is evaluated through different metric categories, as discussed in Section 6.1. Depending on the evaluation objective, variants of ACCURACY, F_1 , RELIABILITY, DISTANCE METRIC, and OTHER (VISUALIZATION or WIN RATE) are used.

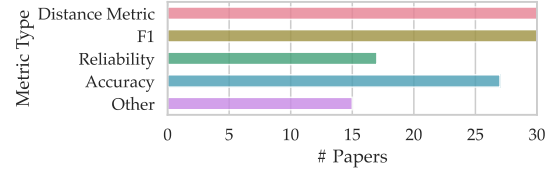
D5.Calibrate the output with respect to frequency of opinions is mainly evaluated via DISTANCE METRICS. Some papers evaluate calibration in terms of model confidence matching empirical performance (Sorensen et al., 2025; Parappan and Henao, 2025). However, such evaluations are not specific to subjectivity.

D6.Represent all perspectives is evaluated from a variety of angles. DISTANCE METRICS are used to compare differences in predictions between demographics or to measure similarity between a model’s default output and specific demographics. RELIABILITY metrics are used to investigate the correlation between the model’s default output and a specific perspective and to measure whether the model can represent demographic groups well. ACCURACY is used for per-country or demographic performance or VALUE INEQUITY INDEX to measure the extent to which model performance deviates across different demographics.

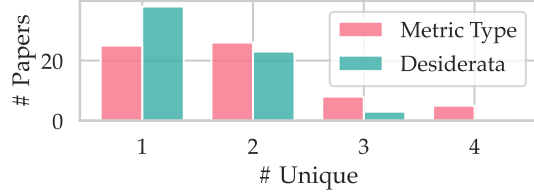
In general, we find that the papers in our analysis focused primarily on D4.Predict and did not investigate D7.Express subjectivity at all (Fig-



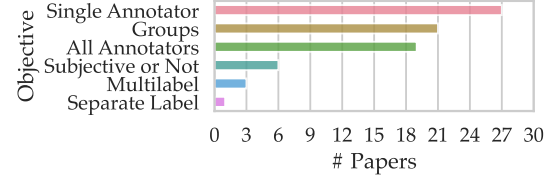
(a) The number of papers that cover each desiderata.



(b) The number of papers that cover each metric type.



(c) The number of unique metric types used and desiderata investigated by the amount of papers.



(d) Number of papers that address a specific evaluation objective.

Figure 4: Statistics from mapping subjective evaluation practices to our desiderata. We examine 64 unique papers, however note that a paper can be counted under multiple metric types and desiderata.

ure 4a). The most prevalent types of metric are DISTANCE METRIC, F_1 , and ACCURACY (Figure 4b). Single annotators are most popular as the evaluation objective, followed by representing groups (Figure 4d). Modeling the full human judgment distribution, not conditioned on a demographic, is also well-studied. Most papers consider one to two metric types or desiderata (Figure 4c).

7 Opportunities and Challenges

We discuss the gaps (🔍) we identified that present opportunities (💡) for future work along with critical challenges, keeping our desiderata in mind.

🔍 The desideratum D7. Express is unaddressed.

None of the papers in our analysis study how to express subjectivity to the user without hampering their experience. Effective modeling of subjectivity is of limited value if the outcome is not adequately conveyed to the user. We can borrow insights and metrics from other fields, e.g., Human-Computer Interaction (HCI) (Ehsan and Riedl, 2020; Benjamin et al., 2022), Affective Computing (Kunc et al., 2024), or Computer-Supported Collaborative Work (Lai et al., 2023), to guide new expression evaluation strategies.

💡 *Use insights and metrics from adjacent fields to evaluate D7. Express.*

🔍 The desiderata D2. Recognize which and D3. Recognize why are understudied. Subjectivity type (D2) and reasons for subjectivity (D3) are addressed implicitly through performance-based

analyses. For D3, only one paper explicitly evaluates disambiguation (in NLI). Some papers resort to qualitative feature-attribution analyses. Interpretability or explainability techniques can play a role in identifying key drivers in the input toward the prediction made. We propose combining these methods with manually annotated data to construct metrics that verify whether the model is making its predictions for the right reasons.

To study D2, we can start from the inherent uncertainty of the input. Traditionally, neural network uncertainty is decomposed into *aleatoric* and *epistemic* uncertainty (Hüllermeier and Waegeman, 2021). Aleatoric uncertainty arises from the input itself, such as ambiguity or noise; measuring it can help identify ambiguous inputs. It remains challenging to measure the uncertainty surrounding the polyphony of an input and we invite future work to tackle this.

💡 *Design metrics to address D2. Recognize which and D3. Recognize why, potentially through methodologies for uncertainty estimation and explainability.*

🔍 The entanglement of polyphony and ambiguity has not been at the forefront yet. These two have been studied in isolation. As illustrated in Section 3, the two types of subjectivity can co-occur and understanding how a model handles this interplay can reveal broader insights into its sensitivity to subjective input.

💡 *Study the entanglement of polyphony and ambiguity at a sample-level.*

🔗 **A multi-faceted and deployment-driven investigation of model behavior is missing.** Considering *all* desiderata is challenging due to limited evaluation practices and experimental constraints. Yet, these aspects should not be studied in isolation but their *interplay* must be studied as well. We find that papers do not evaluate the consequences of earlier desiderata on the subsequent desiderata (e.g., the impact of the model performing badly on D3.Recognize why on D4.Predict), while this is crucial for ensuring reliable model deployment.

💡 **Study the interplay of different desiderata.**

Further, NLP researchers may not know how the model will express subjectivity, particularly in the case of free-text generation scenarios (e.g., overton vs steerable). Current research directions emphasize steering or personalizing a model toward a specific perspective, rather than investigating how the model fares in capturing a wider or even the (full) spectrum of perspectives. Thus, it is essential to evaluate all three operationalizations of *pluralistic alignment* for informed deployment decisions.

💡 **Evaluate modes of expressing subjectivity.**

🔗 **Evaluation of D4.Predict in free-text generation is limited.** Most metrics are directly applicable in classification scenarios, but require additional steps for free-text-generated responses. Future work can focus on extending metrics beyond NLI- and LLM-as-a-judge-based solutions. Using LLMs to evaluate has been found to be unreliable as it can come with biases, such as length, order, or self-preferences (Koo et al., 2024), and can also amplify human-like biases that we would like to mitigate, such as gender or misinformation oversight bias (Chen et al., 2024b).

💡 **Design metrics to evaluate how well free-text generation predicts subjectivity.**

🔗 **Current datasets still fail to meet standards for subjective evaluation.** Datasets are increasingly being released with disaggregated labels and annotator backgrounds (e.g., Sachdeva et al. (2022); Aroyo et al. (2023)), including information such as annotator confidence ratings (Alacam et al., 2025). However, this might not be sufficient to capture the full spectrum of perspectives. Answering questions such as *who* should annotate *what* requires interdisciplinary efforts, grounding annotation strategies in fields such as intersectional feminism (Klumbyèt et al., 2022), moral philosophy (Graham et al., 2013), or social science (Zhao,

2021). Future developments include metrics that can measure the diversity of the annotator distribution for a given sample or datasets with annotations that enable high-quality subjective representations in the model and thorough evaluation approaches.

💡 **Develop interdisciplinary approaches to improve datasets and annotation strategies.**

8 Conclusion

This paper analyzes the current evaluation paradigms of subjectivity-sensitive models. We begin by understanding subjectivity and how it is represented in NLP data and models, which leads us to define **desiderata** for a subjectivity-sensitive model. We take a top-down approach to design *seven* desiderata, driven by our ideals for a model to (1) represent all, based on (2) the correct source of subjectivity to (3) generalize to the real world. We analyze 64 papers to map evaluation metrics to our desiderata. This allows us to identify gaps that can be fruitful directions for future work. Among these gaps, we find that existing work has overlooked nuances in the types of subjectivity: polyphony versus ambiguity, a distinction we argue is crucial for effective evaluation. Additionally, current work does not focus on the desideratum: Express, nor the interplay between different desiderata.

Limitations

Scope of Evaluation Directions This paper aims to demonstrate that the evaluation of subjectivity-sensitive studies remains an open challenge. Our primary objective was to identify key gaps in the evaluation of subjectivity-sensitive models and outline potential directions for future research. Although we discuss potential strategies for evaluating such models throughout the paper, the implementation of concrete evaluation methodologies is left for future work. However, inspiration from the concrete metrics in Section 6.1 can be taken as a starting point.

Literature Survey Methodology To support the discussion of the desiderata introduced in this work, we survey state-of-the-art work in Section 6.2. We emphasize that the goal of this survey is not to provide an exhaustive overview of the field, as done by Frenda et al. (2025); Xu and Jurgens (2026), but rather to obtain a broad characterization of current evaluation practices. The selection of metrics (i.e., annotation labels) for the surveyed papers was

performed in a top-down manner to align with our proposed desiderata. While this choice enables consistent alignment within our paper, it may also have caused certain patterns to be overlooked; for more comprehensive surveys, we refer the reader to [Frenda et al. \(2025\)](#); [Xu and Jurgens \(2026\)](#). Finally, each paper was annotated by a single author, which may introduce some degree of annotation subjectivity. To mitigate this issue, the authors who annotated discussed and calibrated the annotation process after division of papers on a subset of papers (6 samples between one set of two authors and 8 samples between another set of two others, amounting to roughly 10% of the total amount of papers).

Acknowledgments

We thank the anonymous reviewers for their comments that helped improve this paper. All remaining errors are our own. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme, and AlgoSoc, a collaborative 10-year research program on public values in the algorithmic society, both funded by the Dutch Ministry of Education, Culture and Science under the Gravitation programme (project numbers 024.005.017 and 024.004.022) through the Netherlands Organisation for Scientific Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of OCW or those of the AlgoSoc consortium as a whole. The figures and affiliation emojis have been designed using resources from [flaticon.com](#); the chatbot for *Express* and person behind laptop are by Flat Icons, lightbulb + hand by Dewi Sari, and the rest of the emojis by freepik.

References

- Gavin Abercrombie, Tanvi Dinkar, Amanda Ceras Curry, Verena Rieser, and Dirk Hovy. 2025. [Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement](#). In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 63–74, Suzhou, China. Association for Computational Linguistics.
- Özge Alacam, Sanne Hoeken, Andreas Säuberli, Hannes Gröner, Diego Frassinelli, Sina Zarriß, and Barbara Plank. 2025. [Disentangling subjectivity and uncertainty for hate speech annotation and modeling using gaze](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28707–28724, Suzhou, China. Association for Computational Linguistics.
- Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, and 1 others. 2025. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *European Conference on Information Retrieval*, pages 467–478. Springer.
- Richard Alies, Elena Merdjanovska, and Alan Akbik. 2025. [Measuring label ambiguity in subjective tasks using predictive uncertainty estimation](#). In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 21–34, Vienna, Austria. Association for Computational Linguistics.
- Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. [Dices dataset: Diversity in conversational ai evaluation for safety](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc.
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. [Stop measuring calibration when humans disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. [Explanation strategies as an empirical-analytical lens for socio-technical contextualization of machine learning interpretability](#). *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP).
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

- Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, J      Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. [Open problems and fundamental limitations of reinforcement learning from human feedback](#). *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.
- Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024a. [“seeing the big through the small”: Can LLMs approximate human judgment distributions on NLI from a few explanations?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.
- Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025. [PAD: Personalized alignment at decoding-time](#). In *The Thirteenth International Conference on Learning Representations*.
- Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. [“I’ll be there for you”: The one with understanding indirect answers](#). In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. [You are what you annotate: Towards better models through annotator representations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. [Toxicity in chatgpt: Analyzing persona-assigned language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.
- Mark D     , Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. 2022. [Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, page 2342–2351, New York, NY, USA. Association for Computing Machinery.
- Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. [Can LLM be a personalized judge?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.
- Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. [Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement](#). In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management, SAD+ CrowdBias 2018*, pages 11–18. CEUR-WS.
- Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). In *First Conference on Language Modeling*.
- Upol Ehsan and Mark O. Riedl. 2020. [Human-centered explainable ai: Towards a reflective sociotechnical approach](#). In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*, page 449–466, Berlin, Heidelberg. Springer-Verlag.
- Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. [Modular pluralism: Pluralistic alignment via multi-LLM collaboration](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. [When the majority is wrong: Modeling annotator disagreement for subjective tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.
- Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. [Social chemistry 101: Learning to reason about social and moral norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. [Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.
- Paula Fortuna, Juan Soler, and Leo Wanner. 2020. [Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. [Perspectivist approaches to natural language processing: a survey](#). *Language Resources and Evaluation*, 59(2):1719–1746.
- Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane 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 the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188, Miami, Florida, USA. Association for Computational Linguistics.
- 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 the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA. Association for Computing Machinery.
- Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. [The disagreement deconvolution: Bringing machine learning performance metrics in line with reality](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA. Association for Computing Machinery.
- Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. [Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation](#). *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. [Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism](#). In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier.
- Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. [More labels or cases? assessing label variation in natural language inference](#). In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.
- Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. 2025. [Counterfactual reasoning for steerable pluralistic value alignment of large language models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. [Bias Runs Deep: Implicit reasoning biases in persona-assigned LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. [How far can we extract diverse perspectives from large language models?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366, Miami, Florida, USA. Association for Computational Linguistics.
- Sanne Hoeken, Özge Alacam, Dong Nguyen, Massimo Poesio, and Sina Zarriß. 2025. [Not just who or what: Modeling the interaction of linguistic and annotator variation in hateful word interpretation](#). In *Proceedings of the 16th International Conference on Computational Semantics*, pages 63–77, Düsseldorf, Germany. Association for Computational Linguistics.
- Amir Homayounirad, Enrico Liscio, Tong Wang, Catholijn M Jonker, and Luciano Cavalcante Siebert. 2025. [Will annotators disagree? identifying subjectivity in value-laden arguments](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15237–15252, Suzhou, China. Association for Computational Linguistics.
- Eyke Hüllermeier and Willem Waegeman. 2021. [Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods](#). *Machine learning*, 110(3):457–506.
- EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. [Aligning language models to user opinions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.
- Daniil Ignatev, Denis Paperno, and Massimo Poesio. 2025. [Hypernetworks for perspectivist adaptation](#). In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 111–122, Suzhou, China. Association for Computational Linguistics.
- Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. [NUTMEG: Separating signal from noise in annotator disagreement](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2874–2887, Suzhou, China. Association for Computational Linguistics.
- Emily Jamison and Iryna Gurevych. 2015. [Noise or additional information? leveraging crowdsourcing annotation item agreement for natural language tasks](#).

- In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2024. [Personalized soups: Personalized large language model alignment via post-hoc parameter merging](#). In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.
- Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. [Can language models reason about individualistic human values and preferences?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. [Investigating reasons for disagreement in natural language inference](#). *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Uthman Jinadu and Yi Ding. 2024. [Noise correction on subjective datasets](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5385–5395, Bangkok, Thailand. Association for Computational Linguistics.
- Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. [Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks?](#) In *First Conference on Language Modeling*.
- Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. [Hate speech criteria: A modular approach to task-specific hate speech definitions](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Goda Klumbyte, Claude Draude, and Alex S. Taylor. 2022. [Critical tools for machine learning: Working with intersectional critical concepts in machine learning systems design](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, page 1528–1541, New York, NY, USA. Association for Computing Machinery.
- Jan Kocoń, Marcin Gruza, Julita Bielaniec, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE international conference on data mining (ICDM)*, pages 1168–1173. IEEE.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. [Benchmarking cognitive biases in large language models as evaluators](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.
- Dominika Kunc, Joanna Komoszyńska, Stanisław Saganowski, Przemysław Kazienko, Karen S. Quigley, and Lisa Feldman Barrett. 2024. [Embracing subjectivity in affective research: Naturalistic and controlled settings](#). In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 176–180.
- Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2026. [Training and evaluating with human label variation: An empirical study](#). *Computational Linguistics*, 52(1):85–111.
- Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. [Selective explanations: Leveraging human input to align explainable ai](#). *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. [Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.
- Noah Lee, Na Min An, and James Thorne. 2023. [Can large language models capture dissenting human voices?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.
- 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 the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. 2025. [Prefpalette: Personalized preference modeling with latent attributes](#). In *Second Conference on Language Modeling*.
- Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn Jonker, Kyriaki Kalimeri, and Pradeep Kumar Murukannaiah. 2023. [What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric](#). In *Proceedings of the 61st Annual Meeting*

- of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14113–14132, Toronto, Canada. Association for Computational Linguistics.
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. [We’re afraid language models aren’t modeling ambiguity](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.
- Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. 2025. [PERSEVAL: A framework for perspectivist classification evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22334–22359, Suzhou, China. Association for Computational Linguistics.
- Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.
- Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. [The prompt makes the person\(a\): A systematic evaluation of sociodemographic persona prompting for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.
- Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. 2025. [Principled personas: Defining and measuring the intended effects of persona prompting on task performance](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26857–26886, Suzhou, China. Association for Computational Linguistics.
- Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. [Hate personified: Investigating the role of LLMs in content moderation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.
- Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. [Benchmarking distributional alignment of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.
- Erik Miehl, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. 2025. [Evaluating the prompt steerability of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7874–7900, Albuquerque, New Mexico. Association for Computational Linguistics.
- Negar Mokherberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. [Capturing perspectives of crowdsourced annotators in subjective learning tasks](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Benedetta Muscato, Lucia Passaro, Gizem Gezici, and Fosca Giannotti. 2025. [Perspectives in play: A multi-perspective approach for more inclusive nlp systems](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 9827–9835. International Joint Conferences on Artificial Intelligence Organization. AI and Social Good.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. [What can we learn from collective human opinions on natural language inference data?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.
- Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. [Beyond demographics: Fine-tuning large language models to predict individuals’ subjective text perceptions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.
- Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. [The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022.

- Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.
- Pia Pachinger, Janis Goldzycher, Anna M. Planitzer, Julia Neidhardt, and Allan Hanbury. 2025. A disaggregated dataset on English offensiveness containing spans. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 1–14, Suzhou, China. Association for Computational Linguistics.
- Deepak Pandita, Flip Korn, Chris Welty, and Christopher M. Homan. 2026. Forest vs tree: The (n, k) trade-off in reproducible ml evaluation. In *AAAI*, pages 24736–24744.
- Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20322–20338, Suzhou, China. Association for Computational Linguistics.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.
- Dzmitry Pihulski and Jan Kocoń. 2025. Language, culture, and ideology: Personalizing offensiveness detection in political tweets with reasoning llms. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 2146–2155.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Elinor Poole-Dayana, Jiayi Wu, Jiaxin Pei, and Michiel A. Bakker. 2025. Benchmarking overtone pluralism in LLMs. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.
- Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’18/IAAI’18/EAAI’18. AAAI Press.
- Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140,

- Bangkok, Thailand. Association for Computational Linguistics.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. [PersonaGym: Evaluating persona agents and LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.
- Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. [Why don't you do it right? analysing annotators' disagreement in subjective tasks](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Olufunke O. Sarumi, Charles Welch, Daniel Braun, and Jörg Schlötterer. 2025. [The impact of annotator personas on LLM behavior across the perspectivism spectrum](#). In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 121–136, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. [VITAL: A new dataset for benchmarking pluralistic alignment in healthcare](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22954–22974, Vienna, Austria. Association for Computational Linguistics.
- Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. [Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025. [Value profiles for encoding human variation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2047–2095, Suzhou, China. Association for Computational Linguistics.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. [Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. [Measuring and testing dependence by correlation of distances](#). *The Annals of Statistics*, 35(6):2769–2794.
- Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. [Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. [SemEval-2021 task 12: Learning with disagreements](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020.

- A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Michiel Van Der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024a. [Annotator-centric active learning for subjective NLP tasks](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.
- Michiel Van Der Meer, Piek Vossen, Catholijn Jonker, and Pradeep Murukannaiah. 2024b. [An empirical analysis of diversity in argument summarization](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2028–2045, St. Julian’s, Malta. Association for Computational Linguistics.
- Tamara Nikolaevna Vardomskaia. 2018. *Sources of subjectivity*. The University of Chicago.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. [Everyone’s voice matters: quantifying annotation disagreement using demographic information](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’23/IAAI’23/EAAI’23*. AAAI Press.
- Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.
- Xiaolong Wang, Yuanchi Zhang, Ziyue Wang, Yuzhuang Xu, Fuwen Luo, Yile Wang, Peng Li, and Yang Liu. 2025b. [Perspective transition of large language models for solving subjective tasks](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9686–9704, Vienna, Austria. Association for Computational Linguistics.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. [VariErr NLI: Separating annotation error from human label variation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.
- Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. [Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.
- Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. [LLM tropes: Revealing fine-grained values and opinions in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.
- Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. [Detoxifying language models risks marginalizing minority voices](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.
- Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in nlp. *arXiv preprint arXiv:2601.09065*.
- Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2025a. [Diverging preferences: When do annotators disagree and do models know?](#) In *Forty-second International Conference on Machine Learning*.
- Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2025b. [Personalization of large language models: A survey](#). *Transactions on Machine Learning Research*. Survey Certification.
- Kino Zhao. 2021. Sample representation in the social sciences. *Synthese*, 198(10):9097–9115.
- Siyao Zhao, John Dang, and Aditya Grover. 2024. [Group preference optimization: Few-shot alignment of large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Yuxiang Zhou, Hainiu Xu, Desmond C. Ong, Maria Liakata, Petr Slovak, and Yulan He. 2025. [Modeling subjectivity in cognitive appraisal with language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13811–13833, Suzhou, China. Association for Computational Linguistics.

A Literature Review and Annotation

In Table 1 we give an overview of the annotations for the papers we considered. We searched for papers on subjectivity in NLP through snowballing from key papers and search terms such as “subjectivity nlp/llms”, “disagreement nlp/llms”, “perspectivism nlp”, “persona prompting”. We found 64 papers and annotated them as follows: we indicated the category of metrics used (DISTANCE METRIC, F_1 -variant, RELIABILITY metric, ACCURACY-variant, or OTHER), the evaluation objective (single annotator/individual, all annotators/population, specific group (e.g., demographic based), subjective as a separate label, multi-label, or subjective or not) and the covered desiderata.

Inclusion and Exclusion Criteria. We primarily focus on papers explicitly focusing on subjectivity, disagreement, or perspectivism in the field of NLP. For completeness, we also include papers focusing on persona prompting but leave out papers on *personalization*. We do not limit our timeframe but prioritize recent papers. While we aimed to include only peer-reviewed work, we do also include preprints that give novel insights.

Annotation Procedure. Each paper was annotated by one of the authors. To ensure alignment between the authors of the paper who annotated, we re-calibrated our decisions with each other on a handful of samples (6 samples between one of the two authors, and 8 between another set of two others, roughly 10%) to ensure as much consistency as possible and align our understanding of the annotation dimensions. For annotating the desiderata that the paper addresses in its evaluation paradigm, we only consider what the paper explicitly says they evaluate for, not what they implicitly do without their intention (e.g., using distance metrics for whether subjectivity is done correctly but implicitly also checking calibration).

	Desiderata	Reliability	Distance Metric	F ₁	Accuracy	Other	Evaluation Objective
Orlikowski et al. (2025)	D4	✗	✓	✓	✗	✗	Individual
Van Der Meer et al. (2024a)	D4	✗	✓	✓	✗	✗	All, Individual
Wright et al. (2024)	D4, D6	✗	✓	✗	✗	VISUALIZATION	Groups
Feng et al. (2024)	D2, D4	✗	✓	✓	✓	WIN RATE, NLI	Groups, All
Wan et al. (2023)	D1, D3	✗	✓	✓	✗	✗	Subjective or Not
Poole-Dayane et al. (2025)	D4	✗	✓	✗	✓	✗	All
Sorensen et al. (2025)	D4	✗	✓	✗	✓	✗	All, Individual
Jiang and de Marneffe (2022)	D1, D2, D4	✗	✗	✓	✓	✗	Multilabel, Separate Label
Durmus et al. (2024)	D4, D6	✗	✓	✗	✗	✗	Groups
Sun et al. (2025)	D4, D6	✗	✗	✗	✗	DEMOGRAPHIC	Groups
Giorgi et al. (2024)	D4, D6	✓	✗	✗	✗	✗	Groups
Zhou et al. (2025)	D4, D5	✗	✓	✗	✗	✗	All
Dong et al. (2024)	D4	✗	✗	✗	✓	✗	Individual
Orlikowski et al. (2023)	D4	✗	✗	✓	✗	✗	Groups
Santurkar et al. (2023)	D4, D6	✗	✓	✗	✗	CONSISTENCY	Groups
Gupta et al. (2024)	D6	✗	✗	✗	✓	✗	Groups
Wang et al. (2025a)	D3, D4	✗	✓	✗	✗	✗	Groups
Li et al. (2025)	D4	✗	✗	✗	✓	✗	Groups
Homayounirad et al. (2025)	D4	✓	✗	✓	✗	✗	Individual, Subjective or Not
Lo et al. (2025)	D4, D6	✗	✗	✓	✗	✗	All, Groups, Individual
Chen et al. (2025)	D4	✗	✗	✗	✗	WIN RATE	Individual
Gordon et al. (2022)	D4	✗	✓	✗	✓	✗	Individual
Tan and Lee (2025)	D4	✗	✓	✗	✗	WIN RATE	Groups
Samuel et al. (2025)	D4	✗	✗	✗	✗	PERSONAScore	Individual
Miehling et al. (2025)	D4	✗	✓	✗	✗	✗	Individual
Chen et al. (2024a)	D4, D5	✓	✓	✓	✗	VISUALIZATION	All
Lutz et al. (2025)	D4	✗	✓	✗	✗	✗	Groups
Alam et al. (2025)	D1	✗	✗	✓	✗	✗	Subjective or Not
Muscato et al. (2025)	D3, D4	✗	✓	✓	✓	EXPLAINABILITY	All
Sarumi et al. (2025)	D4	✗	✓	✓	✗	✗	Individual
Lee et al. (2023)	D4, D5	✗	✓	✗	✓	✗	All
Liu et al. (2023)	D3, D4	✗	✗	✓	✓	✗	Multilabel
Plepi et al. (2022)	D4, D6	✓	✗	✓	✓	✗	Individual
Zhang et al. (2025a)	D1, D4, D6	✗	✓	✗	✓	✗	Individual
Guo et al. (2025)	D4, D5	✓	✓	✗	✗	✗	Multilabel
Pihulski and Kocoi (2025)	D4	✓	✗	✗	✗	✗	Individual
Meister et al. (2025)	D4, D5	✗	✓	✗	✗	✗	All
Zhao et al. (2024)	D4	✗	✓	✗	✓	✗	Groups, Individual
Luz de Araujo et al. (2025)	D4	✓	✗	✗	✗	✓	Individual
Parappan and Henao (2025)	D4	✓	✗	✓	✗	✓	All, Groups
Ryan et al. (2024)	D4, D6	✓	✓	✓	✓	✗	Groups
Fornaciari et al. (2021)	D4	✗	✗	✓	✓	✗	All
Fleisig et al. (2023)	D4	✗	✓	✗	✗	✗	Groups, Individual
Deng et al. (2023)	D4	✗	✗	✓	✓	✗	Individual
Ignatev et al. (2025)	D4	✓	✗	✓	✗	✗	Individual
Weerasooriya et al. (2023)	D4	✗	✓	✓	✓	✗	All
Kocoi et al. (2021)	D4	✗	✗	✓	✓	✗	Individual
Sandri et al. (2023)	D1, D2, D3	✗	✗	✓	✗	✗	Subjective or Not
Mokherian et al. (2024)	D4	✓	✗	✓	✗	VISUALIZATION	Individual
Alies et al. (2025)	D1, D4	✓	✓	✓	✗	✗	Subjective or Not
Pachinger et al. (2025)	D4	✗	✗	✓	✗	✗	All
Jang et al. (2024)	D4	✗	✗	✗	✗	WIN RATE	Individual
Hoeken et al. (2025)	D3, D4	✓	✗	✗	✓	FEATURE ASSOCIATION	Individual
Kurniawan et al. (2026)	D4	✓	✓	✓	✓	✗	All
Romberg (2022)	D1	✗	✗	✓	✓	✗	Subjective or Not
Lee et al. (2024)	D4, D6	✗	✗	✗	✓	✗	Groups
Masud et al. (2024)	D4, D6	✓	✗	✓	✗	✗	Groups, Individual
Mostafazadeh Davani et al. (2022)	D4	✗	✗	✓	✗	✗	All, Individual
Hwang et al. (2023)	D4	✗	✓	✗	✓	✗	Individual
Jiang et al. (2025)	D4, D6	✗	✗	✗	✓	✗	Individual
Sap et al. (2022)	D6	✓	✗	✗	✗	✗	Groups
Uma et al. (2020)	D4	✓	✓	✗	✓	✗	All
Uma et al. (2021a)	D4	✗	✓	✓	✗	✗	All
Shetty et al. (2025)	D4	✗	✓	✗	✓	WIN RATE, NLI	All, Groups

Table 1: Overview of papers considered and their annotations. **Desiderata**: the desiderata explicitly addressed by the paper, with the following options; D1. Recognize when, D2. Recognize which type, D3. Recognize why, D4. Predict, D5. Calibrate, D6. Represent, and D7. Express. **Reliability**: whether the paper uses a reliability metric. **Distance Metric**: whether the paper uses a distance metric. **F₁**: whether the paper uses F_1 . **Accuracy**: whether the paper uses accuracy. **Other**: whether the paper uses any other metric, if so, which one? **Evaluation objective**: who is the paper evaluating for, with the following options: individual annotator, full human judgment distribution, a specific demographic or group, predicting whether an input is subjective or not, subjective is a separate label to be predicted, and multi-label predictions.