Humans Hallucinate Too: Language Models Identify and Correct Subjective Annotation Errors With Label-in-a-Haystack Prompts

Georgios Chochlakis, Peter Wu, Arjun Bedi,
Marcus Ma, Kristina Lerman, Shrikanth Narayanan
University of Southern California
Correspondence: chochlak@usc.edu

Abstract

Modeling complex subjective tasks in Natural Language Processing, such as recognizing emotion and morality, is considerably challenging due to significant variation in human annotations. This variation often reflects *reasonable* differences in semantic interpretations rather than mere noise, necessitating methods to distinguish between legitimate subjectivity and error. We address this challenge by exploring *label verification* in these contexts using Large Language Models (LLMs). First, we propose a simple In-Context Learning binary filtering baseline that estimates the *reasonableness* of a document-label pair. We then introduce the *Label-in-a-Haystack* setting: the query and its label(s) are included in the demonstrations shown to LLMs, which are prompted to predict the label(s) again, while receiving task-specific instructions (e.g., emotion recognition) rather than label copying. We show how the failure to copy the label(s) to the output of the LLM are task-relevant and informative. Building on this, we propose the **L**abel-**i**n-**a**-**H**aystack **R**ectification (*LiaHR*) framework for subjective label correction: when the model outputs diverge from the reference gold labels, we assign the generated labels to the example instead of discarding it. This approach can be integrated into annotation pipelines to enhance signal-to-noise ratios. Comprehensive analyses, human evaluations, and ecological validity studies verify the utility of *LiaHR* for label correction. Code is available at https://github.com/gchochla/liahr.

1 Introduction

In this work, we address the challenge of modeling complex subjective tasks in natural language, captured in benchmarks such as for emotion recognition and moral foundation prediction. By "complex subjective", we refer to problems where multiple (subjective) interpretations can be *reasonable*, and there is often no single "correct" answer. In such cases, "ground" truth is substituted with crowd truth (Aroyo and Welty, 2015), such as majority vote. Previous work has also referred to these settings as *survey settings* (Resnick et al., 2021), where similarly "ground" truth is the wisdom of the crowd. This stands in contrast to "objective" tasks where we can define a correct answer and annotator disagreement is generally viewed as error or noise. The distinction is evident when looking at inter-annotator agreement in these settings (Mohammad et al., 2018; Demszky et al., 2020), but also the

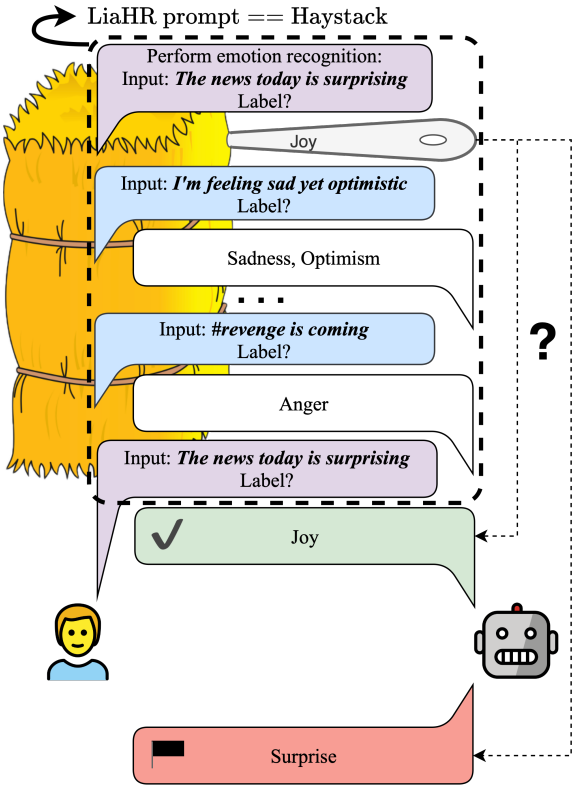

Figure 1: *Label-in-a-Haystack Rectification* (***LiaHR***): The query also appears in the prompt as a demo. The LLM is instructed to perform the actual task, as captured by the label names. We leverage the failure to correctly copy-paste the query's label to flag the query-label pair, for filtering or even correction based on the prediction.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19637–19656
November 4-9, 2025

utility of objectively correct responses compared to disagreements in reinforcement learning with verifiable rewards (Guo et al., 2025), for instance.

Therefore, whereas noise in objective labels needs to be discarded and can be detected by looking at agreement between annotators, for subjective tasks, annotator disagreement may carry signal rather than noise, reflecting differences in perspective or background. Therefore, conventional error correction approaches based on agreement metrics are not directly applicable. Instead, improving subjective modeling requires filtering variation due to error in gold labels while preserving meaningful disagreement (Booth and Narayanan, 2024).

To address this challenge, we propose a framework that uses LLMs for *error detection and correction* in subjective annotations that respects different perspectives. In this manner, we can maintain the diversity of opinions in the data, while also maximizing the signal-to-noise ratio. Prior works in these settings (Hovy et al., 2013; Swayamdipta et al., 2020; Mokhberian et al., 2022) have relied on training classifiers across entire datasets to identify unreliable labels based on model predictions and inter-annotator disagreement. In contrast, our approach leverages LLMs in a few-shot, online setting to assess and even refine labels during annotation. We begin by introducing "reasonableness" labels as the simple baseline (Figure 16 in the Appendix) to demonstrate how LLMs can be catered to filtering explicitly instead of proxy filtering through classification. This binary indicator characterizes whether a document-label pair is reasonable (i.e., plausible, as we do not necessarily adopt a right-wrong split). We can, thereafter, prompt an LLM to predict the reasonableness label of a query document-label pair.

To achieve correction, we introduce the *Label-in-a-Haystack* task, shown in Figure 1, that leverages the biases of LLMs toward their prior knowledge (Shi et al., 2024; Chochlakis et al., 2024). In this setting, the query and a candidate label are included in the prompt, and the model is instructed to perform the task of the dataset (that is, not merely to copy the label). Given the prediction of the LLM, we simply check whether the model was able to copy the labels from its prompt correctly. We refer to this setting as **L**abel-**i**n-**a**-**H**aystack **R**ectification (*LiaHR*), as the model generates alternatives when it "disagrees" enough with the provided labels, effectively correcting unreasonable annotations.

To evaluate our proposed approaches, we first propose, define and evaluate four proxy properties integral to subjective modeling: **Nonconformity**, **Diversity**, **Noise rejection**, and **Rectification**. Then, we verify whether model decisions and proposed alternatives align well with human judgments. Finally, to assess the ecological validity of the filtering and correction proposed, we show that the performance of BERT-based models (Devlin et al., 2019) increases on the corrected datasets.

Our findings reveal that both the reasonableness baseline and the *LiaHR* framework can successfully verify and correct subjective labels. As such, our proposed framework can be effectively used *during* (not after) the annotation process, and is specifically catered to complex subjective settings. We leverage its commonsense priors to correct the labels, rejecting unreasonable annotations in context, reinforcing prior observations that in-context learning in LLMs may rely more on *task recognition* than *task learning* (Min et al., 2022). Furthermore, by causally manipulating the prompt labels to belong to in-group or out-group members (Dorn et al., 2024), but without explicit mention of this manipulation to the model, we show how *LiaHR* can reliably pick up implicit cues from a few examples. Finally, we also show that aggregated labels are rejected at higher rates compared to individual annotators, corroborating previous findings (Chochlakis et al., 2025) of the unsuitability of aggregation for subjective language tasks.

2 Related Work

2.1 Viewpoint Diversity

Many works have attempted to model individual annotator perspectives instead of the aggregate to capture their differing perspectives. Recently, Gordon et al. (2022) fused Transformer features (Vaswani et al., 2017) with annotator features that include demographic factors, among others, to model individual perspectives. Demographic information has also been fused into word embeddings by Garten et al. (2019). In addition, systematic biases have been assessed through rigorous annotations and profiling (Sap et al., 2022). Other recent work has tried to model annotators on top of common representations (Davani et al., 2022; Mokhberian et al., 2023), and to decrease annotation costs online based on disagreement (Golazizian et al., 2024). Modeling annotators with LLMs has shown limited success due to LLM biases (Dutta et al., 2023; Abdurahman et al., 2024; Hartmann et al., 2023).

2.2 Error Detection

Previously, error detection has been carried out in a variety of ways and levels of intervention. One research thread assumes a single correct answer per item, and proceed to identify errors or "spammer" annotators. Examples include the Dawid-Skene algorithm (Dawid and Skene, 1979), MACE (Hovy et al., 2013), and CrowdKit (Ustalov et al., 2021). However, these methods fail the basic assumption of our work, as they do allow difference in opinion, marginalizing idiosyncratic viewpoints, which may otherwise be internally consistent (Chochlakis et al., 2025). Similar approaches that allow for disagreement still assign scores per item and annotator individually and not for separately for each pair, like CrowdTruth (Dumitrache et al., 2018).

In another research thread, again as a post-processing step, previous work has used trained models on the dataset to assess the quality of the labels, either directly, e.g., with dataset cartography (Swayamdipta et al., 2020; Mokhberian et al., 2022), where each data point is mapped onto a 2D space depending on the confidence and the accuracy of the predictions, or indirectly, e.g., with self distillation (Stanton et al., 2021).

Label verification has also been explored online by using predictions from a model, such as a Large Language Model (LLM), and checking them against the annotations (Feng and Narayanan, 2024). However, this method trivially considers differing perspectives invalid. Previous work has also shown how the prior biases of LLMs ossify their posterior predictions (Chochlakis et al., 2024), which in turn leads to failures in accommodating different perspectives during regular inference. This further narrows the breadth of subjective assessment we ideally want to capture and limits our ability to potentially elicit different predictions from LLMs in subjective settings. When iterating in batches, verification checks cannot be automated similar to the aforementioned post-processing step due to the lack of sufficient data, so checks need to be manual, such as analyzing disagreement or having annotators engage in consensus talks (Paletz et al., 2023), significantly increasing costs. Liu et al. (2023) showed that LLMs do not model ambiguity, an important component of disagreement.

3 Methodology

First, it is important to provide some working definition of *reasonableness* (and in turn, what a subjective task is). For our purposes, we consider a document-label pair to be reasonable if and only if a person who would annotate differently can nonetheless consider some reasoning process that leads to that label valid. That is, if a human can agree that a reasoning process is *valid*, *coherent*, and *faithful* (Jacovi and Goldberg, 2020) with respect to the label, then that label is deemed reasonable[1]. We present the general and intuitive description of our methods in this section and a more mathematically rigorous description in Appendix A.

Reasonableness labels We construct a dataset dynamically, wherein our data consist of document-label pairs. As a proxy for reasonableness, the labels are either the gold label of the document from the original dataset, or randomly sampled for unreasonable pairs. This setting is shown in Figure 16 in the Appendix. Each document can appear with both types of labels. We sample the labels of another example from the dataset for unreasonable pairs to maintain the label distribution.

Label-in-a-Haystack As shown in Figure 1, the query and its candidate label are included in the prompt as the first example, and the model is instructed to perform the task described by its labels. However, due to inclusion of the label in the prompt already, we essentially check whether the model is able to copy-paste the query's label onto its output. Given previous results about the collapse of the posterior predictions of LLMs to the priors in complex subjective tasks (Chochlakis et al., 2024), we expect that in cases where the gold labels are "judged" to be erroneous by the model, the copy-pasting will fail, flagging a label for further review. In addition to this ability, this setting also allows us to immediately get alternative labels for the example, a property that the baseline does not possess. In this manner, we do not waste data by discarding examples that are flagged by the model. We note that when using random labels for the query document, we sample them from a random document in the dataset, similar to the baseline.

Intuitively, this method exploits the reliance of the model on its prior knowledge of the task. If a label has sufficiently "high probability" for a model a priori, even if not its dominant prediction, then we expect its presence in the prompt to "push" the posterior towards that label enough so that it prevails

[1] in the case of initial disagreement with a specific rationale, iterative refinement until agreement is achieved is valid, assuming the reasoning remains faithful to the labels

in the output. Therefore, only highly unreasonable labels are rejected by the model, leading to higher precision in identifying errors. Note that the performance of the model using In-Context Learning is rather poor for such tasks (Chochlakis et al., 2024), resulting in poor precision with many false negatives and therefore increased annotation costs.

3.1 Proxy Properties

In this section, we define and present desirable proxy properties that can be used as proxies for the label filtering and correction ability that practitioners can use to guide model selection. Note that since strictness is not required because of their proximate nature, some of them are fuzzy and heuristic.

> **Nonconformity**: The model should flag some dataset labels as unreasonable, but only for a small percentage of examples.

This is the first requirement. Although "small" is nebulous, the model should be copying the gold labels significantly better compared to its performance as a classifier. Having a smaller gap to the dataset's labels indicates an ability to agree with different perspectives, and it assumes that most of the dataset has been annotated properly[2].

> **Diversity**: The model should accept different labels consistently.

Respecting different opinions is also an integral property. Here, we also assume that most annotators have annotated most of the dataset properly[3]. For this quality of the model, we can use annotations from different individuals and expect the model to predict reasonableness or successfully copy the labels at equally high rates for them all.

> **Noise Rejection**: The model should assign reasonableness at random performance levels when using random labels.

That is to say, when asked to "verify" a random label, the model should succeed only when the label "happens" to be reasonable, meaning random levels of performance (though not exactly a random baseline, as more perspectives not present in the data could also be valid). We measure this by randomizing the label of the pair for the baseline or the query label for *LiaHR*, and expect low success rates of filtering or copying respectively.

> **Rectification**: When *LiaHR* is prompted with random labels for the query, its alternative predictions should be closer to the original, gold labels than the random labels it was given.

This final property is a *LiaHR*-specific constraint. If the model is to not only identify unreasonable labels, but also correct them, then when it is given random labels for the query, its predictions should be closer to the gold labels compared to those random labels. As a result, this can be measured by calculating the similarity of the *LiaHR* predictions when *LiaHR* is provided a random query label with the original, gold labels, and comparing that with the copy performance for the random labels, which is equivalent to the similarity of the predictions to the random labels. We expect successful models to have the higher similarity to the gold labels. We expect that priming the model with random labels may cause it to fail to meet this precisely, so only trends towards it are sought out.

3.2 Human Evaluations

To validate our findings on the proxy properties, we perform human evaluations in two settings:

Reasonableness We compare human assessment of the reasonableness of the labels to the LLM's assessments. We use the chi-square test of independence of variables in a contingency table to evaluate the significance of our results (with the binary variable being reasonableness).

Preference We compare human preference for *LiaHR* predictions over the gold label. Significance is calculated with a binomial test. We also compare to the regular ICL predictions to isolate the effects of *LiaHR* from the model's classification capabilities on these tasks.

3.3 Ecological Validity

In addition to human evaluations, we train smaller models on the labels derived from our filtering pipelines. Namely, we examine (i) the `Original` labels, (ii) *LiaHR* on the entire corpus (`Replaced`) or only on the (`trn`) set, (iii) *LiaHR* but used to filter training example when copy-pasting is erroneous

[2] in this assumption, we take for granted that annotators have been screened, trained, attention-checked, etc. Namely, we assume quality data collection

[3] we make the same assumptions as above

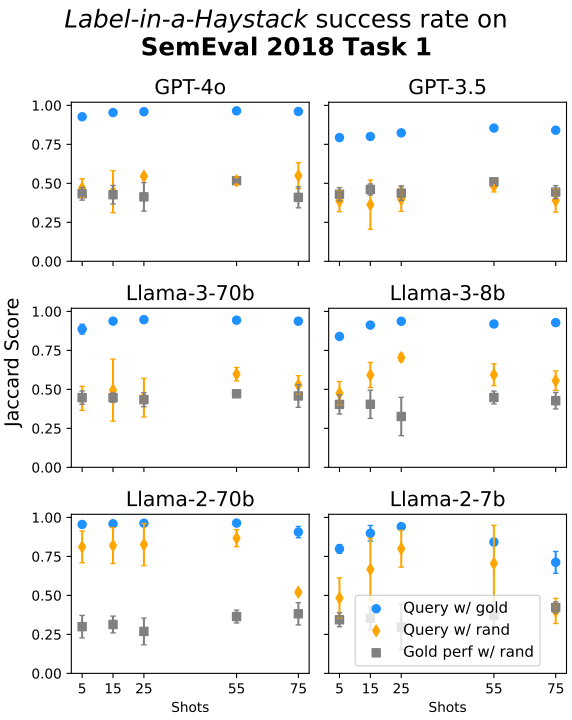

Figure 2: Success rate of copying the labels in *LiaHR* on **SemEval** when using the gold and random labels for the query in the prompt across various numbers of demonstrations. We also show performance w.r.t. the gold labels when using random query labels.

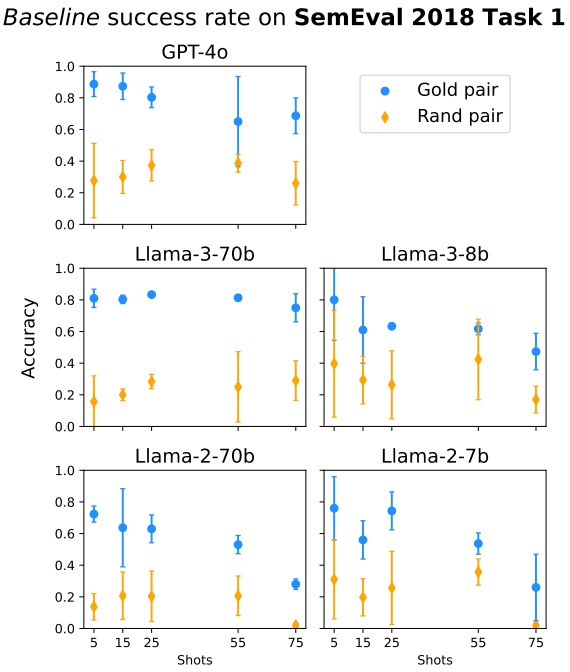

Figure 3: *Baseline* "reasonable" scores on **SemEval** when using gold and random input-label pairs.

(`Filtered`), (iv) the *reasonableness* baseline to filter out training examples (`Bsl Filtered`), (v) the `Predictions` of the LLM with ICL.

3.4 Metrics

Because the *LiaHR* format is identical to classification, we use classification metrics to evaluate the performance of copy-pasting and get a more nuanced picture of the predictions of the model. We use Jaccard Score and Micro F1 for multilabel, and accuracy and F1 for single-label cases.

4 Experiments

4.1 Datasets

SemEval 2018 Task 1 E-c (SemEval; Mohammad et al. 2018) A multilabel emotion recognition benchmark containing annotations for 11 emotions: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *love*, *optimism*, *pessimism*, *sadness*, *surprise*, and *trust*. We use only the English subset.

MFRC (Trager et al., 2022) A multilabel moral foundation corpus with annotations for six moral foundations: *care*, *equality*, *proportionality*, *loyalty*, *authority*, and *purity*. The dataset was released with annotator labels.

GoEmotions (Demszky et al., 2020) A multilabel emotion recognition benchmark with 27 emotions. For efficiency and conciseness, we pool the emotions to the following seven "clusters" using hierarchical clustering: *admiration*, *anger*, *fear*, *joy*, *optimism*, *sadness*, and *surprise*. The dataset was released with annotator labels.

QueerReclaimLex (Dorn et al., 2024) Single-label binary harm dataset, which contains various templates populated with reclaimed LGTBQ+ slurs. It contains two harm labels: assuming in-group and out-group authors. Using one or the other without explicit mention, we can evaluate the **Diversity** property with a known and controllable causal factor. This setting serves as a stress test, since reclaimed slurs in general are a documented failure case for, e.g., toxicity classifiers (Sap et al., 2019; Haimson et al., 2021; Sap et al., 2022), allowing us to examine whether systematic biases in LLMs influence their decisions in our framework. For the same reasons, it is challenging because it includes a realistic confounding factor: the interplay between politeness guardrails and our desired behavior, as slurs are explicitly included throughout the prompt. We create splits to be as balanced as possible, but also present ROC-AUC to avoid bias. Because the labels are binary, we use the opposite label instead

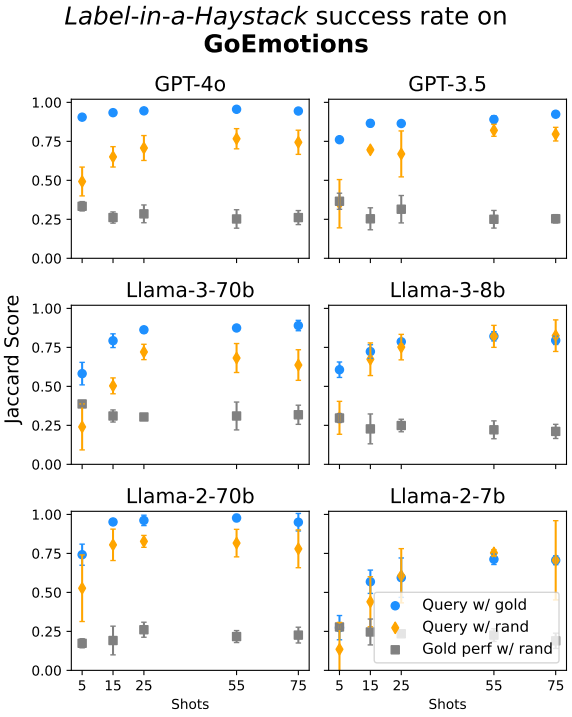

Figure 4: Success rate of copying the labels in *LiaHR* on **GoEmotions** when using the gold and random labels for the query in the prompt across various numbers of demonstrations. We also show performance w.r.t. the gold labels when using random query labels.

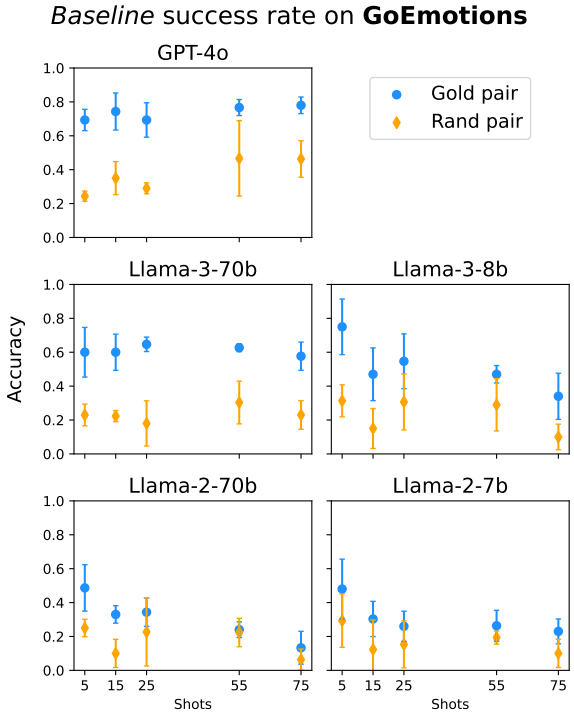

Figure 5: *Baseline* "reasonable" scores on **GoEmotions** when using gold and random input-label pairs.

of randomizing the query label.

4.2 Implementation Details

We use the 4-bit quantized versions of the open-source LLMs through the *HuggingFace* (Wolf et al., 2020) and `bitandbytes` interface for *PyTorch*. We use GPT-3.5 Turbo (`gpt-3.5-turbo`), GPT-4 (`gpt-4-turbo`), and GPT-4o (`gpt-4o-mini`), Llama-2 7B and 70B (`meta-llama/Llama-2-#b-chat-hf`), and Llama-3 8B and 70B (`meta-llama/Meta-Llama-3-#B-Instruct`). We chose only finetuned models (Ouyang et al., 2022) to avoid confounding factors. We use random retrieval of examples. We train *Demux* (Chochlakis et al., 2023) as the smaller model for ecological validity. When sampling random labels, we ensure at least one label is present (i.e., we do not allow `None`s because of their higher plausibility). Results for proxy properties are 3 different seeds with 100 inference examples each. The entire corpus is used for training and evaluation of smaller models. Unless otherwise noted, we show 95% confidence interval around the mean. For more details, see Appendix B and C.

4.3 Evaluating Proxy Properties

The first step to applying these methods for label verification is to show that copy-pasting can fail in *LiaHR*, and that they indeed meet the desired proxy properties. Throughout this section, when presenting **success rates**, that refers to the amount of copy-pasting that happened successfully. This means that *when randomizing* the labels, we still count *whether the random labels are generated*, and therefore *lower scores on random labels* represent more desirable behavior.

SemEval We present our results for all[4] models in Figure 2 for *LiaHR* and Figure 3 for the baseline. In Figure 2, we present the performance of the model on the copy-paste task when using gold (`Query w/ gold`) and random (`Query w/ rand`) labels for the demo query, as well as the performance of the model on the gold labels when the query label is random (and therefore the model has not seen the test label for the query; `Gold perf w/ rand`). All results are shown for 5, 15, 25, 55, and 75 shot (to demonstrate scalability). For Figure 3, we show the first two scenarios, where the docu-

[4]some API-based models were deprecated during the course of our experiments, so we skip them where they are not available. For additional results, such as GPT-4, see Appendix D.

ment is presented to the LLM with its paired label (`Gold pair`) or a random label (`Rand pair`).

In *LiaHR*, we see clear evidence for our desired behavior in bigger and more capable models, specifically GPT-3.5, GPT-4o, and Llama-3 70b. These models seem to display all the properties we check for: **Nonconformity**, **Rectification**, and **Noise rejection**. First, the success rate with gold labels for the query is not perfect (meaning 1.0), yet it is significantly higher compared to the same model's performance on the benchmark (Chochlakis et al., 2024). This means that the model does not conform to the gold labels completely, yet is greatly influenced by them in its predictions (otherwise we would anticipate performance much closer to its "regular" predictions). By meeting both these criteria, the aforementioned models meet the **Nonconformity** property. Then, when we use random labels instead of gold for the query in the prompt, we see the success rate drop dramatically compared to when gold labels are presented (that it, when comparing `Query w/ gold` to `Query w/ rand`). This indicates that models achieve the **Noise Rejection** property. Moreover, it is interesting to see that, when random labels are provided, the predictions match more closely the gold labels (`Gold perf w/ rand`) than these random labels (`Query w/ rand`). Since this criterion is met, the models achieve **Rectification**.

For the "reasonableness" baseline, we see that only GPT-4o meets the criteria **Nonconformity**, and **Noise rejection**[5]. While other models mostly meet the **Noise Rejection** criterion, their success rate is too low to qualify for **Nonconformity**. We also notice that the success rate in all settings is noticeably lower compared to *LiaHR*.

Interestingly, when looking at smaller and less capable models, we see that the models achieve higher copy-paste performance, both with the dataset labels and with random labels, and therefore **Nonconformity** and **Noise Rejection** are only partially achieved. Consequently, when using random query labels, their predictions are more similar to these random labels compared to the dataset labels, so the models do not display **Rectification**.

GoEmotions We show our results for *LiaHR* in Figure 4 and for the baseline in Figure 5. We notice that in GoEmotions, even GPT-4o struggles, with the acceptance rates of random labels, as the gap is smaller to the gold labels when compared to SemEval. Therefore, it is evident that only a small subset of the settings is able to *clearly* achieve **Non-**

[5]**Rectification** is not a potential property because the LLM does not generate labels

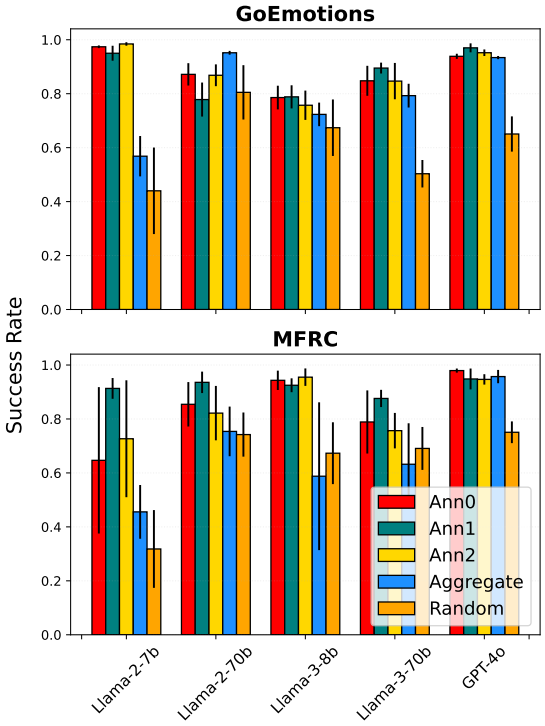

Figure 6: Success rate of copying the labels of *LiaHR* on **GoEmotions** and **MFRC** with aggregate labels, random labels, and annotator labels (*Ann#*), shown for 15-shot prompts. For GoEmotions, actual annotator IDs are: Ann0 = 37, Ann1 = 4, Ann2 = 61. For MFRC: Ann0 = 0, Ann1 = 1, Ann2 = 3.

LiaHR	**Reasonableness**	**Preference**
Llama-3 70b	6.57e-1	**3.36e-2**
GPT-3.5	9.52e-2	7.25e-2
GPT-4	**2.38e-7**	**6.86e-4**
GPT-4o	**1.40e-4**	**5.08e-5**
Baseline		
Llama-3 70b	**6.11e-4**	-
GPT-4o	**8.08e-10**	-
ICL		
GPT-3.5	-	5.19e-1
GPT-4	-	1

Table 1: *p-values* for *LiaHR* on **SemEval**. **Reasonableness** refers to whether human and LLM unreasonable assessments coincide. **Preference** to whether humans prefer model predictions over gold labels. *p-values* are for the hypothesis that the models agree with humans.

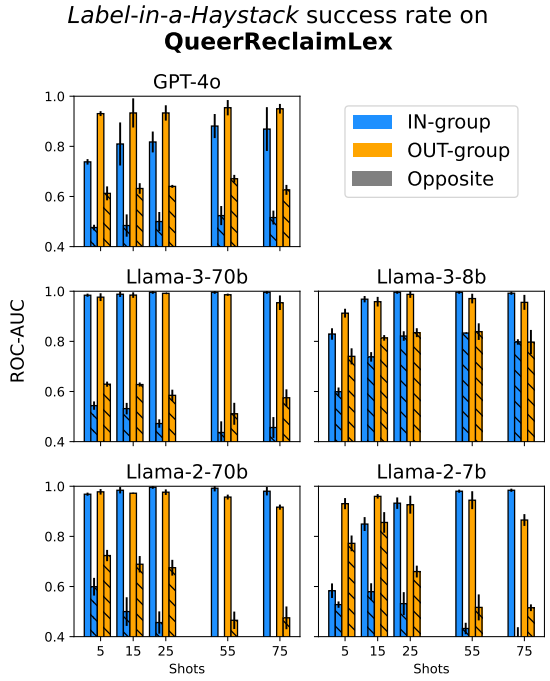

Figure 7: Success rates on copying labels in *LiaHR* on **QueerReclaimLex** when using in-group labels or out-group labels in the prompt as a proxy for **Diversity**. Query included with current group's label or *opposite*.

conformity and **Noise Rejection**, namely 5-shot GPT-4o, 5-shot GPT-3.5, and 15-shot Llama-3 70b, while these models also seem to be meeting or tending towards **Rectification**. Again, the baseline, on the other hand, seems to be achieving consistently lower success rates for the gold labels, but their random performance is much lower and therefore better at **Noise Rejection**.

MFRC In Appendix E, we also show our results and very interesting findings for these three properties in MFRC, where smaller models seem to be treating the gold and random labels similarly.

BERT-based baseline In Appendix L, we show results for a BERT-based filtering baseline, showing it underperforms 5-shot GPT-4o while requiring the entire dataset to be trained, disincentivizing its usage from beginning to end of the data collection.

Diversity We examine **Diversity** separately, in Figures 6 and 7. Figure 6 shows the success rates of copy-pasting on MFRC and GoEmotions between the gold, random, and individual annotator labels, using otherwise the same exact prompts and only differing the labels to avoid confounding factors. We first see that all annotators tend to be clustered together with small rejection rates, indicating that the model tends to accept all different perspectives equally. Second, we see that their performance is better compared to random. Finally, the similarity between the annotators shown can be very low (e.g., as low as 0.433 Jaccard Score on GoEmotions between annotators), representing consistently different perspectives. Consequently, the *majority of the disagreement between annotators is being preserved by the model* organically, without any intervention. All these pieces of evidence indicate that most models achieve **Diversity**. Moreover, we see a marked difference between annotators and the aggregate, with the latter displaying higher rejection rates, indicating that part of our aforementioned results on MFRC and GoEmotions can be explained as aggregation artifacts (Chochlakis et al., 2025).

Setting	Micro F1	
	GoEmotions	SemEval
`Original`	0.652±0.001	0.689±0.002
`Replaced`	**0.653**±0.000	**0.692**±0.003
`Replaced (trn)`	0.642±0.001	0.680±0.002
`Filtered`	0.652±0.002	0.679±0.002
`Bsl Filtered`	0.638±0.001	0.680±0.003
`Predictions`	0.427±0.002	0.613±0.000

Table 2: Performance of BERT-based *Demux* on various settings using *LiaHR* and baseline label corrections.

Figure 7 shows that *LiaHR* can successfully accept both in-group and out-group perspectives in the QueerReclaimLex benchmark without explicit prompting, instead learning implicit causal cues from few examples. Results show that the models tend to model out-group annotations better. However, more capable models also recognize reclaimed slurs as not harmful when used by in-group speakers, scaling performance with more demonstration, indicating the robustness of *LiaHR* to the guardrails placed on models, and an ability to counterbalance systematic biases with few demonstrations, a challenging problem in toxicity classifiers with reclaimed slurs (Sap et al., 2019; Haimson et al., 2021; Sap et al., 2022). Thus, *LiaHR* proves robust to our stress test for whether the model itself might introduce biases in the dataset.

4.4 Human Evaluation

Results for our human evaluations are presented in Table 1 for SemEval for the models that meet our defined properties. More detailed results on **SemEval** and **GoEmotions** can be found in Ap-

pendix C. We see that Llama-3 70b and GPT-3.5 do not show enough discriminability between reasonable and unreasonable labels, although their results are strong in terms of preference for their labels when the copy-paste task was performed incorrectly. However, GPT-4 and 4o can distinguish between reasonable and unreasonable labels and also propose better alternatives for unreasonable labels. The results show strong statistical significance, but also large effect sizes. This is not the case when checking for the ICL prediction of the models. This shows that the predictions of LLMs are not preferred over the gold labels by humans, indicating that our settings are important to achieve proper filtering. We also see that the explicit baseline shows sufficient discriminability for both Llama-3 70b and GPT-4o.

4.5 Ecological Validity

In addition to the human evaluations and defining and evaluating proxy properties, we also perform ecological validity studies, and compare to other online methods. That is, even though we have shown the models have desirable properties, and people tend to prefer them over the original labels, do models trained on them perhaps show erratic behavior? For all the settings introduced in Section 3.3, we show the results in Table 2 (additional results in Appendix H). The results indicate that the new labels lead to slightly better generalization performance, although the methods need to be applied throughout the annotation process to get the maximum benefit. Note that **SemEval** is a smaller dataset, leading to extra performance decreases when examples are filtered instead of corrected. Noticeably, we also see that using the raw predictions of the models leads to substantial deterioration in performance. In addition to the humans evaluations, these results indicate that our proposal for "reasonableness" checks rather than simply using the LLM as classifier is warranted.

5 Conclusion

In this work, we propose "reasonableness" checks to improve the signal-to-noise ratio in subjective language annotations. We leverage LLMs and introduce *LiaHR*, which is able to both filter and correct unreasonable annotations, and a simple baseline that detects unreasonable annotations. We demonstrate that both approaches satisfy desirable proxy properties, pass human evaluations, and show ecological validity when used to train smaller models. Moreover, we show that the model can pick up on causal yet implicit cues from few examples reliably. While our experiments show that humans prefer the model's labels when it is performing correction, we advocate for usage during the annotation process, with additional checks. For example, if some submitted labels for a specific example do not pass the *LiaHR* filter, instead of always using its alternative predictions, the same document can be shown to the annotator at a later stage to verify and potentially correct the label themselves.

To further corroborate our findings on *LiaHR*, we also show how it performs in objective tasks in Appendix F, an analysis of the copy-paste performance across shots, model families and sizes in Appendix G, that individual labels are uniformly affected in Appendix I, and the robustness to the position of the query in Appendix J.

6 Limitations and Ethics

We want to emphasize that our model is not an oracle. The model does not provide ground truth / gold labels and could be biased in other ways.

Our work entails some potential for deliberate misuse. Although we advocate for using individual perspectives as demonstrations in *LiaHR* throughout our work, deliberate misuse might include skewing the perspectives in the prompt and using the rejection from *LiaHR* as justification for rejecting minority labels and preventing certain valid perspectives from entering the data (i.e., gatekeeping). Therefore, we want to emphasize that *LiaHR* assessments can only be considered valid (though *not necessarily correct*) when the perspective being evaluated (the query label) coincides with the perspective in the demonstrations. The predictions of the model should not be taken into account otherwise.

Accidental misuse includes model biases seeping into the labels. We want to note that, despite the remarkable robustness of the framework on the reclaim slurs dataset, **QueerReclaimLex**, its performance on the in-group data is noticeably worse than the out-group. This indicates that there might be some bias in the decisions of the model. Moreover, assessments of harm are inherently subjective, reflecting differences in individual and cultural perceptions. The original work aggregates distinct gender identities (e.g., non-binary, transgender) under the umbrella term gender-queer and treats them

as largely synonymous. While this simplification overlooks the diversity of perspectives, we follow the original work's adoption of this grouping as a pragmatic choice to support our analysis. As understandings of differing perspectives continue to evolve, future work should aim to incorporate a broader pool of annotators and explore methods for capturing the nuance and variability of perceived harm across different communities. Therefore, we urge *immense* caution when the framework is used in sensitive settings.

We also decreased the number of inference queries within each seed to enable us to experiment with many models and shots. This tradeoff means that we do not have a high degree of confidence in each individual result, yet the vast number of experiments demonstrating similar trends reinforces our confidence in our general findings.

A potential confounding factor in our work is quantization. Previous work has reported significant decreases in performance from it (Marchisio et al., 2024). We note, first, that there is no a priori reason for the quantization to affect our results in a nonuniform way, e.g., affecting random labels more than gold labels. Quantization was chosen because of obvious computational constraints. Finally, it is plausible that even API-based models are served quantized (e.g., `mini` versions). For these reasons, we believe that quantized performance is representative of LLM performance in realistic scenarios. Moreover, this work does not aim to establish the benchmark performance of LLMs in any task, but rather to leverage their capabilities to solve a prescient problem in subjective annotations.

Acknowledgments

This project was supported in part by funds from DARPA under contract HR001121C0168, NSF CIVIC, and USC-Capital One Center for Responsible AI Decision Making in Finance. The authors thank Efthymios Tsaprazlis, Efthymios Georgiou, Kleanthis Avramidis and Sabyasachee Baruah for helpful comments.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

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.

Brandon M Booth and Shrikanth S Narayanan. 2024. People make mistakes: Obtaining accurate ground truth from continuous annotations of subjective constructs. *Behavior Research Methods*, 56(8):8784–8800.

Georgios Chochlakis, Gireesh Mahajan, Sabyasachee Baruah, Keith Burghardt, Kristina Lerman, and Shrikanth Narayanan. 2023. Leveraging label correlations in a multi-label setting: A case study in emotion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2024. The strong pull of prior knowledge in large language models and its impact on emotion recognition. In *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2025. Aggregation artifacts in subjective tasks collapse large language models' posteriors. 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 5513–5528, Albuquerque, New 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.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful speech detection by language models exhibits gender-queer dialect bias. In *Proceedings of the 4th ACM Conference on Equity*

and Access in Algorithms, Mechanisms, and Optimization, pages 1–12.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement.

Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. 2023. Modeling subjectivity (by mimicking annotator annotation) in toxic comment identification across diverse communities. *arXiv preprint arXiv:2311.00203*.

Tiantian Feng and Shrikanth Narayanan. 2024. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12116–12120. IEEE.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. *Cognitive science*, 43(1):e12701.

Preni Golazizian, Ali Omrani, Alireza S Ziabari, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. *arXiv preprint arXiv:2402.14101*.

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*, pages 1–19.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *Left-Libertarian Orientation (January 1, 2023)*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual llms? *arXiv preprint arXiv:2407.03211*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Negar Mokhberian, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 147–154. IEEE.

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Susannah BF Paletz, Ewa M Golonka, Nick B Pandža, Grace Stanton, David Ryan, Nikki Adams, C Anton Rytting, Egle E Murauskaite, Cody Buntain, Michael A Johns, et al. 2023. Social media emotions annotation guide (SMEmo): Development and

initial validity. *Behavior Research Methods*, pages 1–51.

Paul Resnick, Yuqing Kong, Grant Schoenebeck, and Tim Weninger. 2021. Survey equivalence: A procedure for measuring classifier accuracy against human labels. *arXiv preprint arXiv:2106.01254*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

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. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34:6906–6919.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2021. Learning from crowds with crowd-kit. *arXiv preprint arXiv:2109.08584*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \ Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

A Methodology

In this section, we present a mathematical formulation for the baseline and *LiaHR* to avoid any potential ambiguities arising from the natural language description of the main text. We follow the notation of Chochlakis et al. (2025). For a set of examples $\mathcal{X}$, and a set of labels $\mathcal{Y}$, a dataset $\mathcal{D}^a$ defines a mapping $f^a : \mathcal{X} \rightarrow \mathcal{Y}$, where a denotes a specific annotator or the aggregate. Similarly, $\mathcal{D}^a = \{(x, y) \mid x \in \mathcal{X}, y = f^a(x)\}$, which is characterized by joint distribution $p^a(x, y)$. The *gold* query pair is denoted as $(x_q, y_q^a),\ y_q^a = f^a(x_q)$.

Reasonableness Baseline We want to sample k train documents from $\mathcal{X}$ to create a prompt with document-label pairs, as well as corresponding binary reasonableness labels, denoted simply as 1 and 0[6]. We choose half ($\frac{k}{2}$) pairs to have the "reasonable" label, and for other half the "unreasonable" label. To sample reasonable pairs for our prompt, we sample document-label pairs directly from $\mathcal{D}^a$ as $S^r = \{(x_i, y_i, 1) :\ (x_i, y_i) \sim p^a,\ i \in [\frac{k}{2}]\}$. For unreasonable pairs, we sample the documents x and the labels y independently of each other from the dataset as $S^u = \{(x_i, y_i, 0) :\ (x_i, y_i) \sim p_I^a,\ i \in [\frac{k}{2}]\}$, where $p_I^a(x, y) = p^a(x)p^a(y)$, in effect assigning random yet in-distribution labels to each document. The complete demonstrations for the model are $S = S^r \cup S^u$, and the order of the examples in the prompt is determined randomly. The query document is presented with the gold label $y_d^a = y_q^a$ (like S^r; "Gold pair" in the baseline figures like Figure 3) or a random label y_d^a independently from the query x_q using $p^a(y)$ (like S^u; "Rand pair" in the baseline figures) at the end, and we elicit the final prediction from the model, namely 1 or 0: $S' = S \cup \{(x_q, y_d^a, \_)\}$. We show the rate of 1 predictions in our results.

LiaHR To create the prompt, we sample k demonstrations for the prompt from $\mathcal{D}^a$ with p^a as $S = \{(x_i, y_i) :\ (x_i, y_i) \sim p^a,\ i \in [k]\}$. In *LiaHR*, the first demonstration is the query x_q. For this demonstration, we either use its gold label $y_d^a = y_q^a$ ("Query w/ gold" in the *LiaHR* figures like Figure 2) or sample a random label y_d^a independently from the query x_q using $p^a(y)$ ("Query w/ rand" in the *LiaHR* figures). This query pair is included in the prompt as the first demonstration, and the query document x_q is appended in the prompt, eliciting a prediction for it from the model, $S' = \{(x_q, y_d^a)\} \cup S \cup \{(x_q, \_)\}$. We measure and show the similarity between the predictions of the model with y_d^a, and only measure the similarity of the prediction with y_q^a for "Gold perf w/ rand" in *LiaHR* figures.

[6] in the actual prompt, the labels are presented as "yes" or "no", or "reasonable" and "unreasonable", not as 1 and 0, as shown in Table 3.

B More Implementation Details

We used A100 NVIDIA GPUs with 80GB VRAM for 70B models, and A40 NVIDIA GPUs for smaller models. The budget for OpenAI API calls was less than $50.

For all datasets, we evaluate LLMs on the dev set. For QueerReclaimLex, we only maintain the labels with agreement between the two annotators. Our splits in the dataset were random. The evaluation set was balanced, containing 84 examples.

For the baseline, we sample the random labels for the pair similarly to the random labels in *LiaHR*. In the demonstrations, we use equal amounts of pairs with gold labels and random labels. For Demux, we use the same training regime as Chochlakis et al. (2023), using the intra loss with a coefficient of 0.2, but training only on the train set instead of integrating the dev set in training after early stopping. Confidence intervals are shown from 3 model runs.

We present examples of all the prompts in Table 3. Across each dataset, the same examples are used in the prompt within each seed when the number of shots is equal. That means that, for example, in QueerReclaimLex, the only factor that is different across the four settings presented in Figure 7 is the causally controlled labels reflecting in-group or out-group perspectives.

C Full Human Evaluations

In human evaluations, to avoid biasing annotators towards specific answers — for example, having the dataset label always as the first option (Turpin et al., 2024) —, we randomly interleave reasonable and unreasonable examples (label according to the model) for the **Reasonableness** check, whereas for the **Preference** check, we randomly change the order with which the label in the dataset and the model's prediction are presented to the annotators. We present full results, including the number of trials and the precise numbers of them, here in Table 4 and Table 5. We note that we recruited 11 colleagues (6 males and 5 females, ages 20-

SemEval 2018 Task 1 E-c

Classify the following inputs into none, one, or multiple the following emotions per input: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise and trust.

Input: ‘Skipping meals is the national way of showing disappointment and . wiz‘
{"label": ["anger", "disgust", "sadness"]}

Input: ‘But guess what ? I’m sober‘

GoEmotions

Classify the following inputs into none, one, or multiple the following emotions per input: joy, optimism, admiration, surprise, fear, sadness and anger.

Input: ‘Is this in New Orleans?? I really feel like this is New Orleans.‘
{"label": []}

Input: ‘BANG motherfucker¡

MFRC

Classify the following inputs into none, one, or multiple the following moral foundations per input: authority, care, equality, loyalty, proportionality and purity.

Input: ‘Le Pen was absolutely terrible. That’s pretty much all we can take out of this debate.‘
{"label": []}

Input: ‘It’s just capitalism is a pure distillation of those bad traits‘

QueerReclaimLex

Consider whether the following inputs present harm or not to the mentioned LGBTQ group, and answer with: no harm and harm.

Input: ‘fags ew‘
Assessment: no harm

Input: ‘went from edm junkie to college dyke way too quick‘
Assessment:

Reasonableness baseline

Assess the reasonableness of the provided label for each input. Namely, evaluate whether the label makes sense for its corresponding input, under some reasonable interpretation. Reply only with unreasonable and reasonable.

Input: ‘Skipping meals is the national way of showing disappointment and . wiz‘
Label: surprise, optimism
Assessment: unreasonable

Input: ‘But guess what ? I’m sober‘
Label: joy
Assessment:

Table 3: Prompt template examples

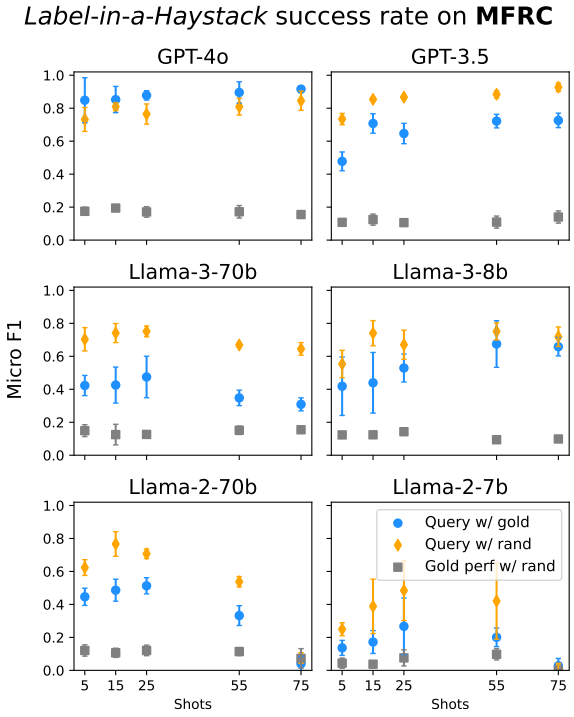

Figure 8: Success rate of copying with *LiaHR* on **MFRC** when using the gold and random labels for the query in the prompt across various numbers of demonstrations. We also show performance w.r.t. the gold labels when using random query labels.

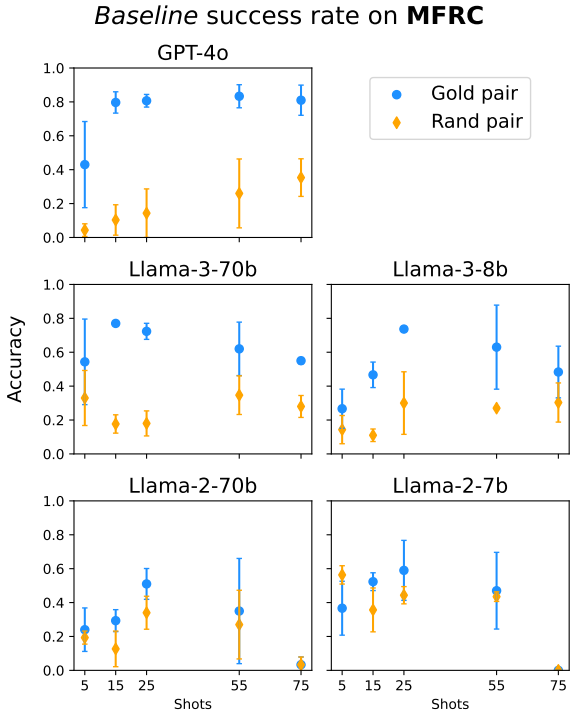

Figure 9: *Baseline* “reasonable” scores on **MFRC** when using gold and random input-label pairs.

28, students or researchers) to annotate to get as many perspectives as possible and avoid biasing the result. Note that the annotators were shown

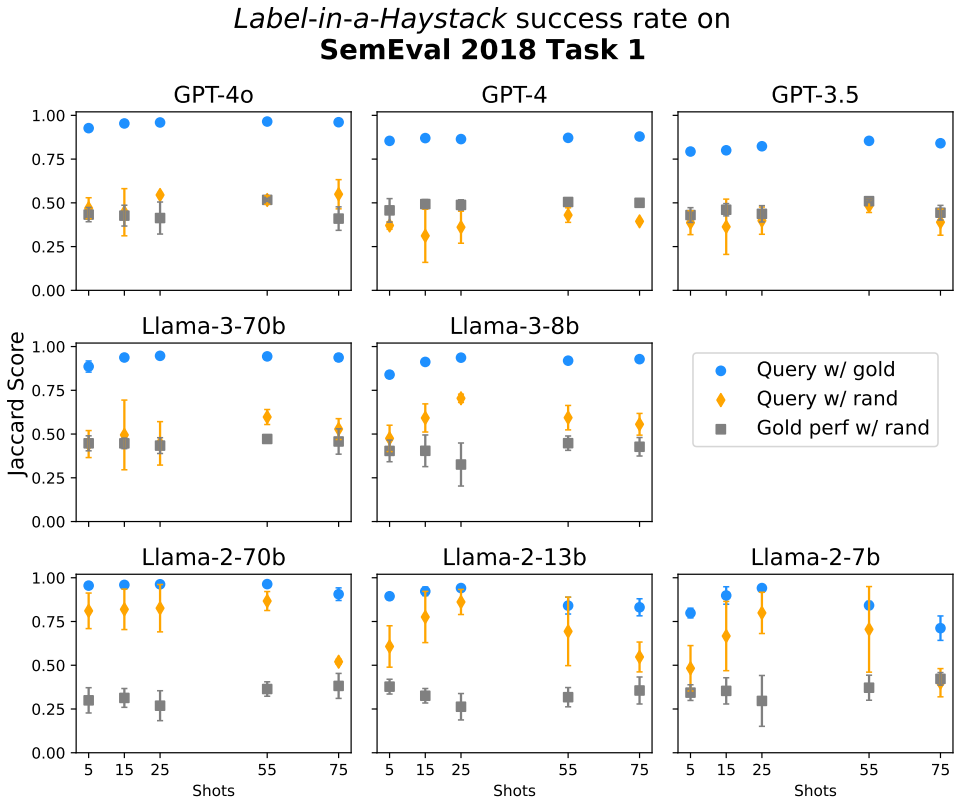

Figure 10: *Full* scores on the copy-paste task on **SemEval** when using the gold and random labels for the query in the prompt across various numbers of demonstrations. We also show performance w.r.t. the gold labels when using random query labels.

the *Reasonableness baseline* prompt from Table 3, modified appropriately.

D More models on SemEval properties

Here, we present additional results on SemEval with some deprecated models present in Figure 10. We see, interestingly, that GPT-4 shows a better performance profile than GPT-4o, indicating that the models have successfully been trained to become more compliant to the user, even if the model disagrees, potentially decreasing the utility of *LiaHR*.

E MFRC properties

In this section, we present the results for **Non-conformity**, **Rectification**, and **Noise rejection** in MFRC, in Figures 8 and 9.

We observe that even GPT-3.5 does not achieve **Noise Rejection** and **Rectification**, but GPT-4o is showing positive trends in the criteria we have. Interestingly, there seem to be settings were random labels perform better than the gold ones. Here, we hypothesize that this happens because we always sample at least one label for the random label case, whereas the dataset contains many examples with no labels.

F Results on objective tasks

Here, we present some experimental results on an objective task, the **T**ext **RE**trieval **C**onference (**TREC**) question classification benchmark (Li and Roth, 2002; Hovy et al., 2001), which contains annotations for the type of information the question pertains to, and specifically Abbreviation, Entity, Description and abstract concept, Human being, Location, and Numeric value. We show these results to verify the intuition that, in principle, *LiaHR* can be used for objective tasks too. Indeed, we see in Figure 11, the system meets our defined properties, with the **Rectification** being, in fact, very strong in this objective setting, suggesting the models in some ways, at least implicitly, learn to represent the nuanced difference between objective and subjective tasks.

		Reasonableness			Preference	
		Correct Ratio	Wrong Ratio	p-value	Ratio	p-value
SemEval	*LiaHR*					
	Llama-3 70b	31/14	28/17	6.57e-1	26/12	3.36e-2
	GPT-3.5	27/8	29/16	9.52e-2	54/36	7.25e-2
	GPT-4	25/5	4/26	2.38e-7	41/15	6.86e-4
	GPT-4o	60/20	21/31	1.40e-4	49/16	5.08e-5
	baseline					
	Llama-3 70b	48/12	29/31	6.11e-4	-	-
	GPT-4o	90/10	49/51	8.08e-10	-	-
GoEmotions	*LiaHR*					
	GPT-4o	43/14	9/27	5.12e-6	36/3	3.61e-8
	baseline					
	GPT-4o	57/23	33/47	2.47e-4	-	-

Table 4: Results of statistical ananlysis for *LiaHR* on **SemEval** and **GoEmotions**. **Correct Ratio** refers to proportion of dataset labels deemed reasonable vs. unreasonable by annotators when the model performed the copy-paste task correctly, and similarly for **Wrong Ratio** when the copy-paste task was performed incorrectly. **Ratio** reflects the times the model's labels were preferred over the gold labels (when the model performed copy-pasting incorrectly).

Model	Preference	
	Ratio	p-value
GPT-3.5	33/27	0.519
GPT-4	28/32	1

Table 5: Results of statistical analysis for the regular ICL / raw predictions setting on **SemEval**. **Ratio** reflects the times the model's predictions were preferred over the gold labels.

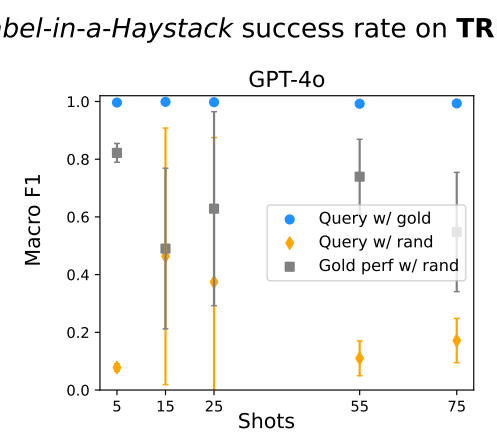

Figure 11: Scores with *LiaHR* on **TREC** (objective benchmark) when using the gold and random labels for the query in the prompt across various numbers of demonstrations. We also show performance w.r.t. the gold labels when using random query labels.

G Degradation in copy-paste performance

In this section, as a summary of our results, we present how different model families and scale affects the drop in copy-paste performance when switching from the gold label for the demo query to a random label in *Label in a Haystack*. We demonstrate the results for SemEval in Figure 12, for GoEmotions in Figure 15, and MFRC in Figure 13. It is interesting to look at the three model families and observe that the more capable the model family is, the larger the degradation in performance tends to be. Moreover, within each family, the larger models usually end up with worse degradation, except for the least capable Llama-2 in some instances, where the trend is the opposite. We therefore hypothesize that there is a U-shaped trend, where, on the lower end, the ability to better follow instructions leads to smaller degradations in performance when shifting to random labels. However, as models continue to get larger, the pull of the priors on the posteriors becomes greater (Chochlakis et al., 2024), leading to greater degradation.

H Extra Ecological Validity results

For completeness, we also present the Jaccard Score for our ecological validity studies to supplement the Micro F1 present in the main body.

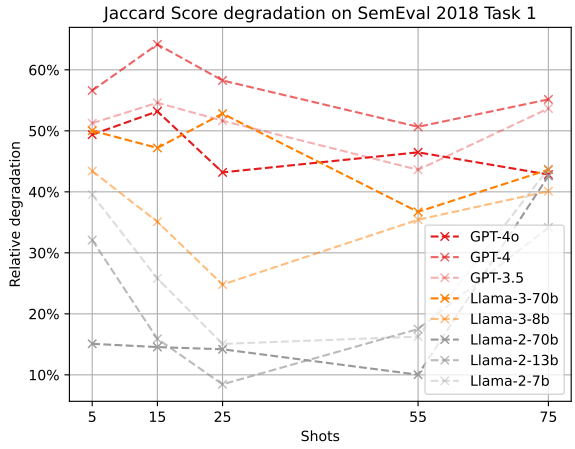

Figure 12: Degradation in copy-paste performance on **SemEval** when using random labels compared to the dataset's labels.

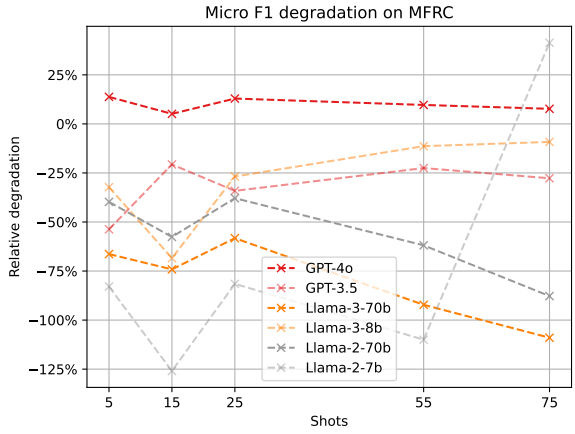

Figure 13: Degradation in copy-paste performance on **MFRC** when using random labels compared to the dataset's labels.

Setting	Jaccard Score	
	GoEmotions	SemEval
`Original`	0.623±0.001	**0.574**±0.001
`Replaced`	**0.624**±0.002	**0.574**±0.003
`Replaced (trn)`	0.615±0.001	0.562±0.001
`Filtered`	**0.624**±0.003	0.561±0.002
`Bsl Filtered`	0.615±0.002	0.558±0.002
`Predictions`	0.430±0.004	0.474±0.000

Table 6: Performance of BERT-based *Demux* on various settings using LLM label corrections.

Results in Table 6 show similar as in Table 2.

I Filtering per Label

We present the success rate of each individual label for our 3 main datasets in Table 7, 8, and 9 based on a 25-shot run with GPT-4o. We see that no label is disproportionately affected, except *trust* in SemEval, the label with the least amount of annotations. On GoEmotions, scores are generally lower compared to GoEmotions, reflecting the clustering process that has been applied to shrink the label set to a reasonable amount.

Emotion	F1
anger	0.972±0.016
anticipation	0.921±0.017
disgust	0.939±0.019
fear	0.977±0.016
joy	0.965±0.010
love	0.973±0.019
optimism	0.995±0.007
pessimism	0.922±0.034
sadness	0.994±0.008
surprise	1.000±0.000
trust	0.867±0.094

Table 7: Success rates of *LiaHR* on SemEval using 25-shot GPT-4o.

Emotion	F1
admiration	0.950±0.021
anger	0.973±0.000
fear	1.000±0.000
joy	0.871±0.020
optimism	0.908±0.036
sadness	0.930±0.028
surprise	0.944±0.020

Table 8: Success rates of *LiaHR* on GoEmotions using 25-shot GPT-4o.

Moral foundation	F1
authority	0.889±0.157
care	0.939±0.043
equality	0.978±0.031
loyalty	0.974±0.036
proportionality	1.000±0.000

Table 9: Success rates of *LiaHR* on GoEmotions using 25-shot GPT-4o.

J Position of Label in the Haystack

We also experiment with changing the position of the query in the prompt and evaluating how all our metrics change. We present our results in Figure 14,

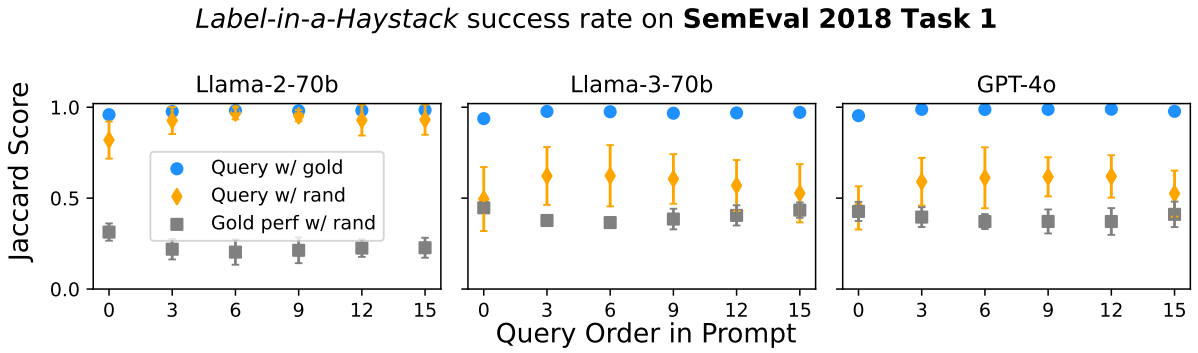

Figure 14: Scores on the 15-shot *LiaHR* on **SemEval** when changing the position of the query in the demonstrations.

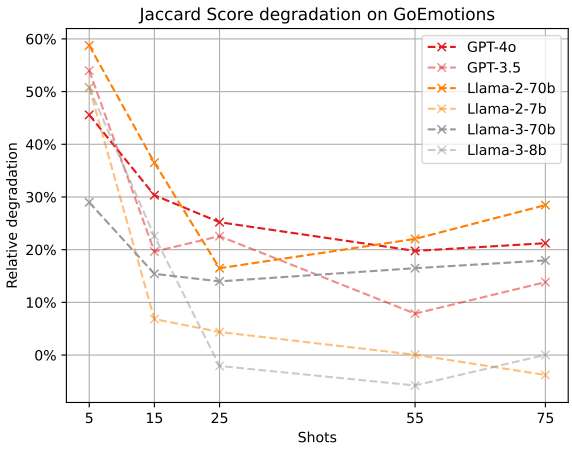

Figure 15: Degradation in copy-paste performance on **GoEmotions** when using random labels compared to the dataset's labels.

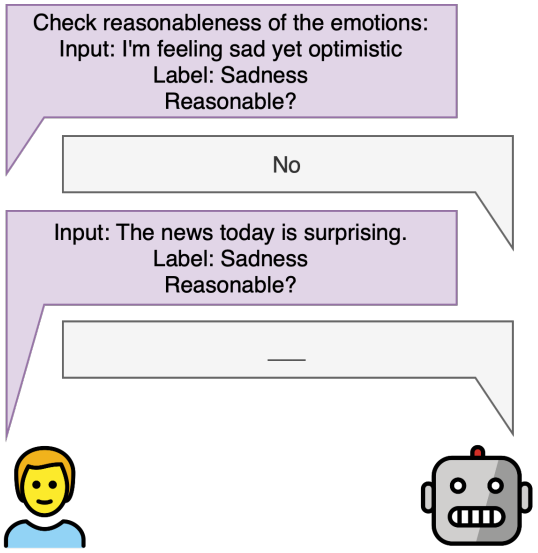

Figure 16: Reasonableness labels: The model is instructed to perform a reasonableness check, as captured by the label names. However, we check for the ability of the model to correctly copy-paste the query's label from its prompt.

with standard deviations shown. We see that no major changes are observed in the predictions of the model, irrespective of where the query appears in the demonstrations. It is very interesting to see that even when the query is the last demonstration (just before itself then), the results remain remarkably similar to when it appears first in the prompt, separated by 15 examples with itself.

K Overall Reasonableness of Annotations

We can estimate the overall reasonableness of the datasets by using our existing analyses. For example, we present the derivation process for SemEval using the GPT-4o (15-shot) *LiaHR* results. First, looking at Figure 2, we can derive the percentage of human annotations predicted to be reasonable by *LiaHR*, $p(\textit{LiaHR}\text{ reasonable}) = 0.954$. Then, focusing on Table 4, we can derive the proportion of examples annotated as unreasonable by our annotators both when *LiaHR* predicted reasonable and unreasonable, namely $p(\text{reasonable} \mid \textit{LiaHR}\text{ reasonable}) = \frac{42}{60}$, $p(\text{reasonable} \mid \textit{LiaHR}\text{ unreasonable}) = \frac{12}{39}$. Finally, we can estimate the overall reasonableness of the annotations as:

$$\begin{aligned} &p(\text{reasonable}) = \\ &p(\text{reasonable} \mid \textit{LiaHR}\text{ reasonable}) \\ &\qquad \cdot p(\textit{LiaHR}\text{ reasonable}) \\ &+p(\text{reasonable} \mid \textit{LiaHR}\text{ unreasonable}) \\ &\qquad \cdot (1 - p(\textit{LiaHR}\text{ reasonable})) \\ &\qquad = 0.682. \end{aligned} \tag{1}$$

The same estimate, when checking with Llama-3 70b, comes to 0.625, and with GPT-3.5 to 0.727. The results are only an approximation, since *LiaHR* results are presented in Jaccard Score, not accuracy.

The same procedure can be used with the baseline, deriving more theoretically sound estimates.

L BERT-based Filtering Baseline

In this section, we present results for a BERT-based filtering baseline, *FilterBERT*. *FilterBERT*'s output is the binary decision of whether to filter a document-label pair out of the data pool. It is a proxy supervised baseline, as we do not use actual annotated data for reasonableness, but instead use the same strategy as for the LLM baseline to construct data. Namely, documents that are paired with their gold label from the dataset are considered "reasonable" pairs. To create "unreasonable" pairs for a document, we sample the labels from a random document in the dataset. Practically, we use all pairs from the original dataset, and for each document we also create an "unreasonable" pair, doubling the size of the dataset. The input is formatted similarly to Demux (Chochlakis et al., 2023), where the input consists of the `CLS` token, followed by the candidate labels, in turn followed by a `SEP` token, and finally the input document. An example input, therefore, is "`[CLS] anger, anticipation, optimism [SEP] I DIDN'T ASK FOR THIS EITHER IT JUST HAPPENED`". We use the contextual embedding of `CLS` with a two-layer neural network (again, similar to Demux) to make the final binary prediction with a threshold of 0.5 on the output sigmoid function. Training details are otherwise identical to Demux (note that we have removed the intro loss coefficient because we do not apply the classifier on each emotion of the prompt). Our results are presented in Table 10, in comparison to the 5-shot GPT-4o baseline results we have already presented in Figures 3, 5 and 9.

The BERT-based model cannot be used to conduct the ecological validity tests due to the fact that it itself needs to use the train and dev sets to be trained, so a more direct comparison is not possible with our current setting, a shortcoming of using a BERT-based model for filtering. However, from these existing results, GPT-4o seems to align more with our intuitions of how a model should perform. For SemEval, the performance of GPT-4o is closer to random baseline performance compared to BERT in rejecting emotions, and accepts more labels. For GoEmotions, the BERT-based model seems to learn the noise in GoEmotion arising from the hierarchical clustering that we apply, achieving higher acceptance rates (as we have mentioned before). The superiority of GPT-4o on MFRC is evident.

	SemEval	GoEmotions	MFRC
FilterBERT			
Gold pairs	0.824 ± 0.017	0.751 ± 0.010	0.181 ± 0.006
Rand pairs	0.244 ± 0.010	0.215 ± 0.011	0.039 ± 0.008
GPT-4o (5-shot)			
Gold pairs	0.887 ± 0.078	0.693 ± 0.063	0.430 ± 0.254
Rand pairs	0.277 ± 0.235	0.243 ± 0.029	0.043 ± 0.038

Table 10: Filtering accuracy for "reasonable" vs. "unreasonable" label-document pairs using a proxy supervised BERT-based classifier (trained on all data) and GPT-4o (5-shot). Gold pairs match the true label; Rand pairs use randomly sampled labels.