Aggregation Artifacts in Subjective Tasks Collapse Large Language Models' Posteriors

Georgios Chochlakis[1], **Alexandros Potamianos**[1,2], **Kristina Lerman**[1], **Shrikanth Narayanan**[1]
[1]University of Southern California, [2]National Technical University of Athens
Correspondence: chochlak@usc.edu

Abstract

In-context Learning (ICL) has become the primary method for performing natural language tasks with Large Language Models (LLMs). The knowledge acquired during pre-training is crucial for this *few-shot* capability, providing the model with *task priors*. However, recent studies have shown that ICL predominantly relies on retrieving task priors rather than "learning" to perform tasks. This limitation is particularly evident in complex subjective domains such as emotion and morality, where priors significantly influence posterior predictions. In this work, we examine whether this is the result of the aggregation used in corresponding datasets, where trying to combine low-agreement, disparate annotations might lead to annotation artifacts that create detrimental noise in the prompt. Moreover, we evaluate the posterior bias towards certain annotators by grounding our study in appropriate, quantitative measures of LLM priors. Our results indicate that aggregation is a *confounding factor* in the modeling of subjective tasks, and advocate focusing on modeling individuals instead. However, aggregation does not explain the entire gap between ICL and the state of the art, meaning other factors in such tasks also account for the observed phenomena. Finally, by rigorously studying annotator-level labels, we find that it is possible for minority annotators to both better align with LLMs and have their perspectives further amplified.[1]

1 Introduction

Large Language Models (LLMs) (Radford et al., 2019; Ouyang et al., 2022a; Touvron et al., 2023; Dubey et al., 2024; Brown et al., 2020; Achiam et al., 2023) have become to dominate language processing as generalists that can perform many tasks. This dominance comes from the emergence of methods such as In-Context Learning (ICL; Brown et al. 2020) and Chain-of-Throught prompting (CoT; Wei et al. 2022), wherein LLMs perform tasks by leveraging input-output demonstrations and task instructions in the prompt only, without any parameter updates.

While ICL is often contrasted with traditional in-weights learning (i.e., gradient-based updates of the models' parameters) (Kossen et al., 2023; Chan et al., 2022), the ICL abilities of LLMs depend on their general, in-weights prior knowledge, allowing them to perform many tasks in a zero-shot or few-shot manner. Therefore, studying how LLMs leverage the context in relation with their existing knowledge is a prerequisite to understanding ICL.

Prior work found evidence that LLMs may be overly reliant on their prior knowledge, disregarding the demonstrations in the prompt. Specifically, Min et al. (2022) demonstrated that, under certain circumstances, LLMs ignore the provided signal in their prompt in the form of the mapping between inputs and outputs, and instead act as a database of tasks; they focus on the examples and the labels *independently* to fetch the underlying task (Xie et al., 2021): Min et al. (2022) sampled examples and labels independently, and showed very little change in performance. Since no annotations are provided, the setting is given the status of "zero-shot" inference, namely **task-recognition zero-shot**. While follow-up work (Kossen et al., 2023; Wei et al., 2023; Pan et al., 2023; Yoo et al., 2022) has further studied this phenomenon and questioned its generality, more recent work (Chochlakis et al., 2024, 2025) has provided further, quantitative evidence for the relative importance of prior and evidence in the posteriors of the models. Specifically, in *complex subjective* tasks like multilabel emotion or morality recognition, LLMs seem to virtually disregard evidence from the dataset's mapping in their posterior predictions, even in the form of Chain-of-Thought prompting (CoT; Wei et al. 2022), performing significantly worse than traditional algo-

[1]code and data available at: https://github.com/gchochla/aggregation-artifacts-llms

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
April 29 - May 4, 2025

rithms (Alhuzali and Ananiadou, 2021; Chochlakis et al., 2023). This may imply an inability to perform annotator modeling, as the same document can receive different but valid annotations by different people.

Here, we use *complex subjective* to denote *survey settings* (Resnick et al., 2021), in which people can reasonably disagree about their semantic interpretations, where the notion of ground truth is replaced with *crowd* truth (Aroyo and Welty, 2015).

In this work, we question whether we can efficiently use this crowd truth when modeling these tasks with LLMs. We hypothesize that the aggregation process creates artifacts in the labels provided to the model, and in turn in the prompt, causing the model to ignore the entire input-label mapping as noise. Specifically, aggregation can create *inconsistent* annotations, since, for instance, different annotators can prevail in different examples, which can cause inconsistencies within or across splits (for a toy example, consult Section C). We create a carefully crafted experimental setting to test our hypothesis with annotator-level labels using ICL and CoT, and try to gauge at the magnitude of this effect. We find that LLMs do indeed tend to consistently favor individual perspectives compared to the aggregate, and in fact favor minority annotators more than majority annotators, who better resemble the aggregate. Our contributions are:

- We show strong correlational evidence that aggregation creates artifacts that hinder the modeling of subjective tasks with LLMs
- We show that minority annotators can both align with LLMs' priors better and demonstrate larger positive effects in the posterior.
- We nonetheless conclude that there are more major factors hurting LLM performance in machine learning benchmarks.

We advocate for more transparency in data collection and data sharing, and urge releasing and modeling individual annotations and not simply the aggregates in benchmark datasets.

2 Related Work

2.1 In-Context Learning and Priors

ICL (Brown et al., 2020) has been used extensively to evaluate LLMs on standard benchmarks (Srivastava et al., 2022). It requires no gradient-based interventions, which are otherwise costly to perform for large models, and usually achieves competitive or state-of-the-art performance. The existence of commercial APIs (Achiam et al., 2023), coupled with open-weights alternatives (Touvron et al., 2023; Dubey et al., 2024) have made ICL an accessible generalist for language tasks and more (Liu et al., 2024). Previous work has further studied controlling the reliance on context and in-weights knowledge through the distribution of the training data (Chan et al., 2022), examining how to optimally select examples for the prompt (Rubin et al., 2022; Gupta et al., 2023), integrating instructions explicitly during training (Touvron et al., 2023; Ouyang et al., 2022b), etc. Relevant to our work, prior work has elicited ICL priors by providing random labels for the examples of the prompt (Min et al., 2022). The resulting minimal variations in performance indicates that LLMs recognize and retrieve their prior knowledge of the task in the prompt rather than doing any "learning". Subsequent results challenged the view that LLMs mostly perform task recognition, showcasing a significant degradation in performance when scaling the prompt (Kossen et al., 2023) or when focusing on specific tasks instead of aggregates (Yoo et al., 2022), and analyzed behavior with certain label manipulations (Kossen et al., 2023; Pan et al., 2023; Wei et al., 2023). More recent work, however, suggests that in complex subjective tasks like emotion and morality classification, the prior understanding of the task dominates posterior predictions (Chochlakis et al., 2024).

One potential way to augment ICL and overcome the prior bias is with CoT (Wei et al., 2022). CoT incorporates the derivation process from input to output explicitly in the prompt, presenting a more human-like reasoning process. This has several advantages, such as making some patterns in the data explicit in the prompt, making model responses more explainable, and potentially directing more computing resources towards more complex problems. However, detailed analysis has cast some doubt on the reliability and the faithfulness of this reasoning technique (Lanham et al., 2023; Turpin et al., 2024). Nonetheless, CoT seems to improve the robustness and performance of LLMs across a plethora of tasks, yet follow-up work on subjective tasks showed no improvements and the same bias towards a *reasoning prior*, especially for larger models (Chochlakis et al., 2025). Methods such as Tree of Thoughts (Yao et al., 2024) or self-consistency (Wang et al., 2022) have experimented with ways to further augment CoT. In this work, we study whether part of the prior bias in both ICL and CoT can be explained by the aggregation process used to derive the labels, and in turn the reasoning process, of the examined benchmarks.

2.2 Annotator Disagreement and Modeling

Many works have attempted to model individual annotator perspectives instead of the aggregate, like we advocate in this work. For example, researchers used the EM algorithm (Dawid and Skene, 1979) to assign confidence to each annotator's evaluations (Hovy et al., 2013). Recently, Gordon et al. (2022) concatenated features derived from Transformers (Vaswani et al., 2017) with annotator embeddings that incorporate demographics to model individual perspectives. Demographic information has also been incorporated in word embeddings by Garten et al. (2019). Demographic information and psychological profiles have been statistically examined in text annotations to derive insights into systematic biases (Sap et al., 2022). Recent work has tried to filter annotators based on deep learning methods (Mokhberian et al., 2022), to model annotators on top of common representations (Davani et al., 2022; Mokhberian et al., 2023), and to decrease annotation costs based on agreement (Golazizian et al., 2024). Modeling annotators with LLMs has shown limited success, and LLM biases have also been explored (Dutta et al., 2023; Abdurahman et al., 2024; Hartmann et al., 2023).

3 Methodology

We closely follow the methodology and 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, as well as reasoning chains $R^a(x) = r, x \in \mathcal{X}$ that explicitly describe f^a, and therefore $\mathcal{D}^a = \{(x, y, r) : x \in \mathcal{X}, y = f^a(x), r = R^a(x)\}$, from which we can sample demonstrations with $p(x, y, r)$. We do not differentiate between splits for brevity. Given CoT prompt $S = \{(x_i, y_i, r_i) : (x_i, y_i, r_i) \sim p_S, i \in [k]\}$ with k demonstrations sampled with distribution p_S from $\mathcal{D}^a$ without replacement, an LLM produces its own mapping and predictions for the task, denoted as $\hat{f}_k(.; p_S) : \mathcal{X} \rightarrow \mathcal{Y}$. When using regular ICL, we simply drop the reasoning text. For all our experiments, we set the temperature

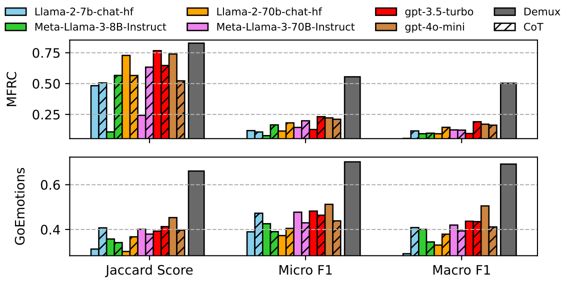

Figure 1: Performance comparison between LLMs w/ and w/o Chain-of-Thought prompting compared to Demux (BERT-based) using aggregated labels.

$\tau = 0$ to derive deterministic predictions, and vary the examples and/or labels between runs.

3.1 Similarity and Performance Metrics

To keep evaluations consistent when using API-based LLMs, we rely on similarity measures calculated directly on the final predictions rather than probabilistic measures like logits. Therefore, we use the Jaccard Score (JS), Micro and Macro F1 metrics (Mohammad et al., 2018) to evaluate the performance of the models. For consistency, we also use them to quantify the similarity between the predictions from different model runs or annotators, since they are symmetric functions which we apply to interchangeable sets.

3.2 Task Prior Knowledge Proxies via Zero-shot Inference

Here, we precisely define the priors that we use for ICL and CoT. We use the term **prior** to contrast it with the **posterior** predictions of the model after *evidence* (i.e., a specific annotator's labels) from the dataset have been presented to it. First, we have the true **reasoning task-recognition zero-shot**[2] **prior**, where the prompt contains k demonstrations sampled with $p^I(x, y, r) = p(x)p(y)p(r)$, so text, labels, and reasoning are sampled *independently* from each other from $\mathcal{D}^a$, hence labels and reasoning are irrelevant to the text and each other. This effectively maintains the relationships between labels, which are strong in such multilabel tasks (Cowen and Keltner, 2017). For ICL, we have the corresponding **task-recognition zero-shot prior** sampled with $p^I(x, y) = p(x)p(y)$, so text and labels are also sampled independently. The similarity of the priors to annotators and aggregate are, therefore, measured by comparing

[2] since no annotations are required

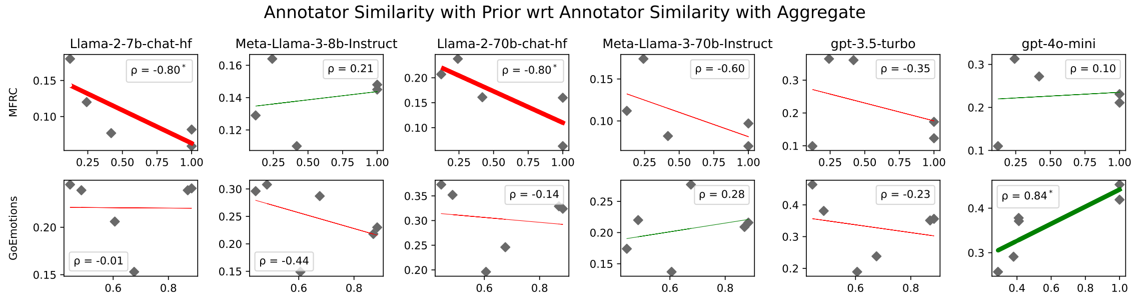

Figure 2: Scatter plot of annotator similarity with aggregate and with prior. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

(as described in Section 3.1) the prior predictions $\hat{f}_k(.; p_I)$ with the annotator's labels, which is equivalent to the prior performance for the annotator, and the posteriors are simply defined as the $\hat{f}_k(.; p_S)$ with the joint sampling distribution p_S (meaning we present the gold labels to the model).

3.3 Prompt Design

Because the specific examples and their order in the prompt can affect the output of the model, we use exactly the same examples and in the same order across corresponding experiments. To achieve that, we find groups of annotators with significant overlap in the train and in the evaluation sets and use the same samples, including for aggregate and prior. Since the only degree of freedom is the labels, we eliminate the aforementioned confounding factors, among others (more details in Section D).

4 Experiments

4.1 Datasets

MFRC (Trager et al., 2022): Multilabel moral foundation corpus with annotations for six moral foundations: *care*, *equality*, *proportionality*, *loyalty*, *authority*, and *purity*. We use annotators `00` through `04` (common examples between groups `00-01-03` and `02-04`).

GoEmotions (Demszky et al., 2020): Multilabel emotion recognition benchmark with 27 emotions. For efficiency and conciseness, we pool the emotions to the following seven "clusters" by using hierarchical clustering: *admiration*, *anger*, *fear*, *joy*, *optimism*, *sadness*, and *surprise*. We use annotator triplets `4-37-61` and `7-36-60`.

4.2 Implementation Details

We use the 4-bit quantized versions of the open-source LLMs through the *HuggingFace* (Wolf et al., 2020) interface for *PyTorch*. We use LLaMA-2 7B and 70B, LLaMA-3 8B and 70B, GPT-3.5 Turbo, and GPT-4o mini. We chose only models with RLHF (Ouyang et al., 2022b) for uniformity. We perform 3 runs for each LLM experiment, varying the examples used. Statistical significance is calculated with permutation tests and measured by considering all 3 runs as separate data points. We use random retrieval of examples. We use less shots for CoT given the increases in length in the prompt. We generated reasonings for each example per annotator and for the aggregate. For details on our CoT annotations, see Section B in the Appendix. We use one NVIDIA A100 and one V100.

4.3 Baselines

To establish baseline performance of LLMs compared to smaller, gradient-based methods, we present performance with and without CoT prompting compared to BERT-based (Devlin et al., 2018) Demux (Chochlakis et al., 2023). In Figure 1, we demonstrate the significant difference in performance across all the LLMs (45 shots for ICL, 15 shot for CoT) and Demux. In fact, given the very high JS and very low F1 scores, results for MFRC indicate close to random performance for the model, so we choose to use Micro F1 for MFRC, as opposed to JS for GoEmotions.

Nevertheless, we argue that this is an artifact of the inconsistent mapping used in the prompt, caused by the aggregation of labels for different annotators. Next, we evaluate whether aggregation does create annotation artifacts, and the extend to which they influence model behavior.

4.4 Main Results

In this section, we present our experiments, aimed at disentangling the role of aggregation in subjective tasks. First, we focus on 45-shot ICL (Section 4.4.1), and analyze how the similarity of each

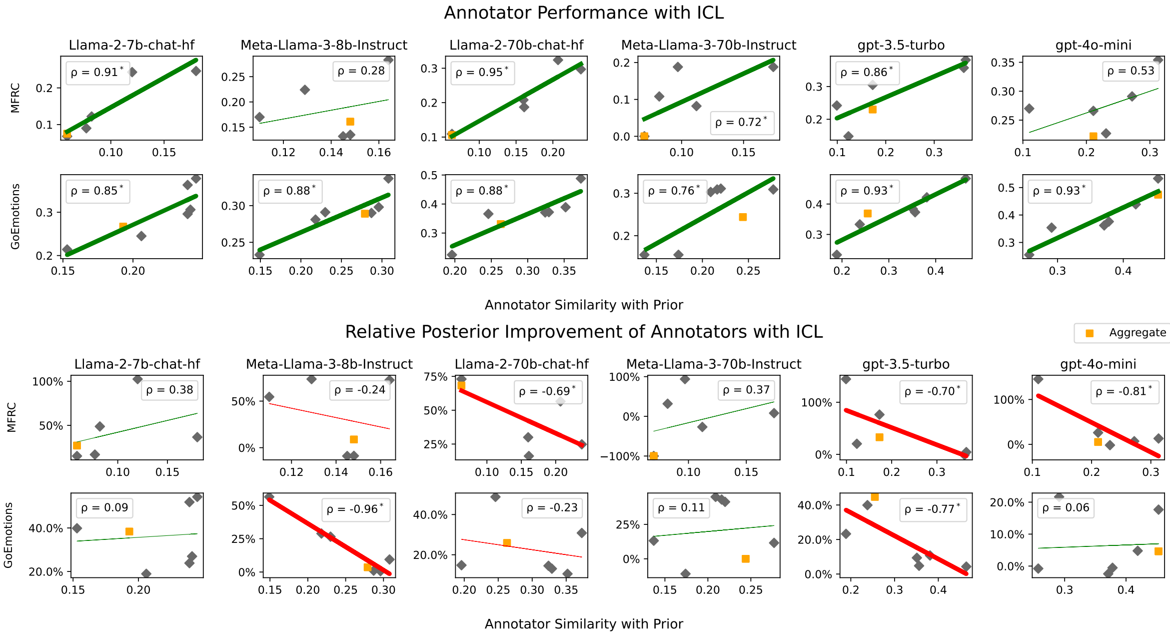

Figure 3: **In-Context Learning** performance for annotators and aggregate and their relative improvement compared to the model's prior as a function of the similarity of each with the model's prior. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

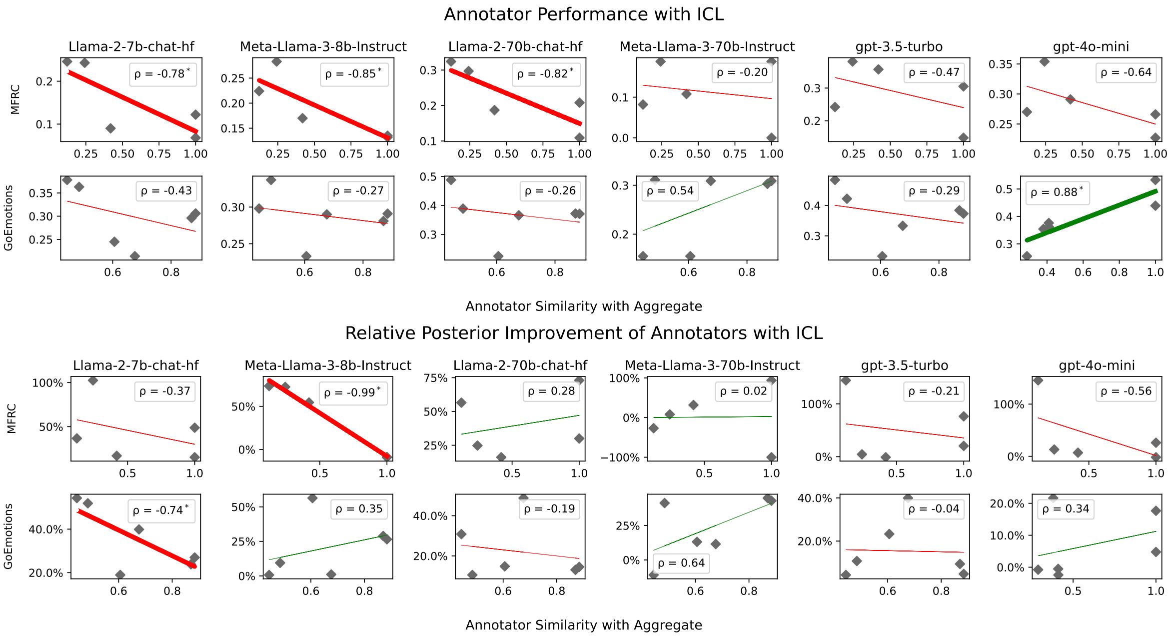

Figure 4: **In-Context Learning** performance for annotators and their relative improvement compared to the model's prior as a function of the similarity of each annotator with the aggregate. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

annotator and of the aggregate with the models' prior affects the relative improvement of the posterior p_S over the prior p_I, as well as absolute posterior performance. Then, we analyze how the majority and minority (or idiosyncratic) annotators fare on these tasks by looking at the similarity of each annotator to the aggregate. Following this analysis, we perform equivalent experiments for 15-shot CoT, and present them in Section 4.4.2. Finally, we zoom out of individual experiments and summarize our complete body of evidence in Section 5.

In all experiments, we present averages. The similarity of the annotators to the prior, as well as the similarity to the aggregate, is measured on the

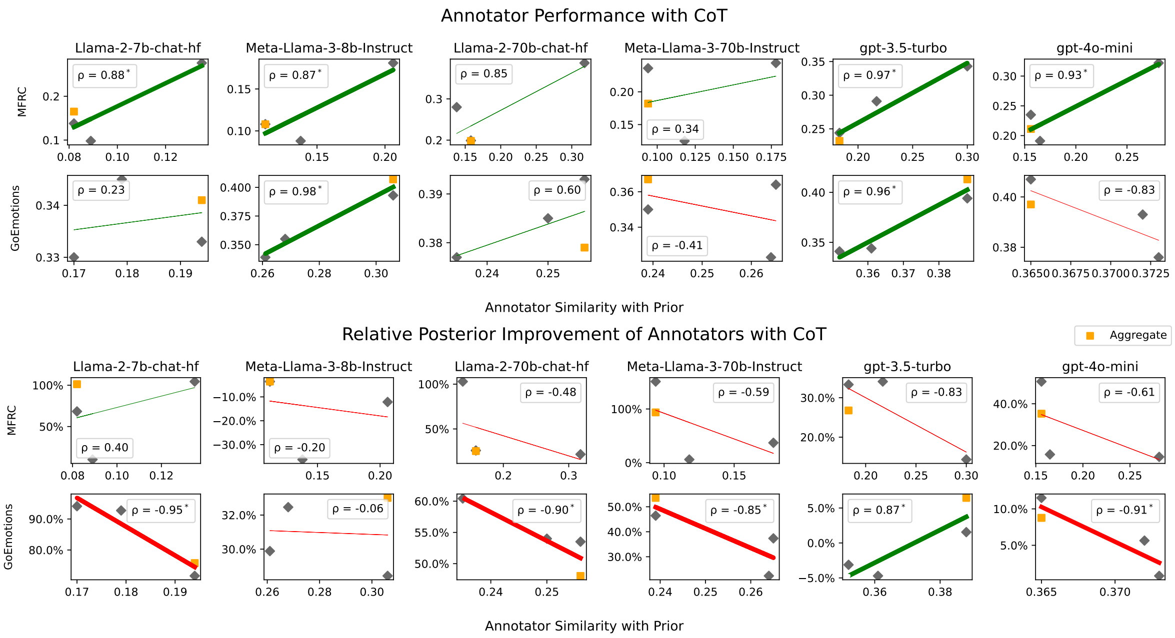

Figure 5: **Chain-of-Thought** performance for annotators and aggregate and their relative improvement compared to the model's prior as a function of the similarity of each with the model's prior. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

test set, which may differ from the train set.

First of all, we present the relationship between the similarity of each annotator to the aggregate and the similarity to the prior of the model. Our results in Figure 2 indicate that the correlation tends to be negative, with 7 of 12 negative correlations, 2 of which are statistically significant, and only 4 positive correlations with one statistically significant result. This indicates that the more an annotator resembles the majority, the less aligned they tend to be with the models' priors. This is the first piece of evidence indicating that aggregation causes artifacts, and seems to suggest that the models perform more appropriate aggregation of evidence during their training compared to the simple (or even simplistic) majority-based aggregation used in such benchmarks, causing misalignment.

4.4.1 Annotator Modeling with ICL

Similarity to Prior In Figure 3, we first see the performance of the models for each annotator and the aggregate w.r.t. the similarity of each to the prior of each model. For the performance of the model, we see that, as expected, similarity with the prior correlates positively with final performance, with all results but one being statistically significant. It is interesting to see that the aggregate ranks low both for MFRC and GoEmotions in terms of posterior (from left to right and top to bottom: 5/6, 4/6, 6/6, 6/6, 5/6, 6/6, 5/7, 5/7, 6/7, 5/7, 5/7, 2/7; average is 22nd percentile) and prior (5/6, 2/6, 6/6, 6/6, 4/6, 5/6, 6/7, 4/7, 5/7, 2/7, 5/7, 2/7; average is 33rd percentile). Looking at the relative improvement, it is interesting to see that the only significant trends are negative trends, meaning that the LLMs tend to boost opinions they disagree with more. Despite the aggregate being among the worst performing mappings, with the expectation being that it receives significant gains in performance, it ranked below average (4/6, 4/6, 2/6, 6/6, 3/6, 5/6, 4/7, 5/7, 3/7, 6/7, 1/7, 4/7; average is 39th percentile).

Agreement with Aggregate By switching to examining at the similarity of each annotator to the aggregate, and how that correlates with absolute and relative performance, in Figure 4, we see strongly negative trends. In fact, 17 out of the 24 cases are negative, 6 of which are statistically significant, and only one positive trend is statistically significant. Therefore, we see that idiosyncratic annotators both have better performance and are more amplified.

Overall, we see very strong correlational evidence that not only are aggregates misaligned with the models' priors, they also benefit less from ICL with their labels in the prompt. This is happening in spite of annotators with worse alignment frequently receiving significant gains in performance. Consequently, by combining our findings from the priors and the aggregates, we conclude that the aggregate

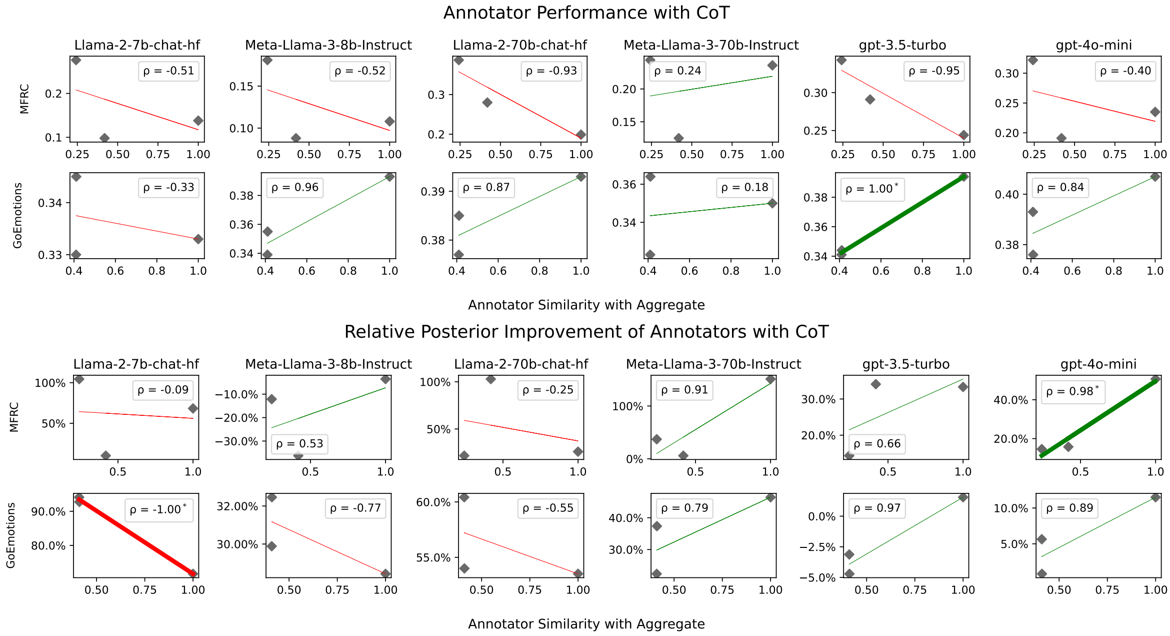

Figure 6: **Chain-of-Thought** performance for annotators and their relative improvement compared to the model's prior as a function of the similarity of each annotator with the aggregate. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

mapping inherently has inconsistencies that inject sufficient noise in the prompt.

4.4.2 Annotator Modeling with CoT

We now switch to CoT, and evaluate consistency across prompting techniques. We note that because of the decreased number of runs in this setting, the confidence in these findings is similarly decreased.

Similarity to Prior In Figure 5, we observe similar trends to the equivalent setting for ICL (Figure 3), namely that final (posterior) performance positively correlates with the prior similarity (10 of 12 cases are positive, 6 of which are statistically significant), and that relative improvement of the posterior compared to the prior is negatively correlated with similarity with the prior (10 of 12 cases are negative, 4 of which are statistically significant). That being said, the differences in performance in Figure 5 tend to be smaller than in Figure 3. The aggregate is among the worst performers in MFRC, but the results in GoEmotions are equivocal.

Agreement with Aggregate In Figure 6, we present CoT results and correlate them with the similarity to the aggregate. Our experiments here seem to be split between negative and positive trends.

Here, due to the decreased number of experiments, it is difficult to extract concrete findings, yet we can ascertain that the findings here do not seem to contradict our previous findings, and do no show improved performance compared to ICL modeling.

4.5 Detailed Analysis

Based on observations during our manual annotation efforts, we identified clear patterns that annotator `01` in MFRC provides the *Authority* label more frequently even when one authority figure is mentioned in the input, in contrast to the rest of the annotators. Therefore, if any learning is achieved from ICL and CoT, we expect the accuracy for *Authority* to be visibly improved compared to the prior baseline due to the consistency and the clear pattern. To achieve and test that, we made this implicit bias clear in the generated reasoning chains. We present the change in performance in Table 1 in comparison to another label, *Equality*, chosen at random, since we did not observe any other clear patterns in other labels. We do indeed observe that the gains in *Authority* F1 score are consistent and tend to be significant across models and settings (with only GPT-4o mini CoT presenting an insignificant drop), in opposition to *Equality*, where gains tend to be small and insignificant, and large drops in performance are observed. This does indicate an ability for the models to somewhat learn and revise priors from the prompt *when the mapping and/or the rationale presented are consistent and clear*.

Model	Authority **F1**		Equality **F1**	
	Prior	Post	Prior	Post
`Llama-2-7b-chat-hf` ICL	0.135±0.121	↑ 0.217±0.153	0.204±0.059	↓ 0.133±0.189
`Llama-2-7b-chat-hf` CoT	0.157±0.124	⇑ 0.330±0.030	0.524±0.140	⇓ 0.197±0.004
`Meta-Llama-3-8B-Instruct` ICL	0.115±0.083	⇑ 0.276±0.033	0.039±0.055	↑ 0.074±0.105
`Meta-Llama-3-8B-Instruct` CoT	0.082±0.116	⇑ 0.277±0.044	0.300±0.216	↑ 0.481±0.086
`Llama-2-70b-chat-hf` ICL	0.035±0.050	↑ 0.053±0.075	0.080±0.063	↑ 0.133±0.189
`Llama-2-70b-chat-hf` CoT	0.107±0.097	⇑ 0.257±0.053	0.044±0.063	⇑ 0.250±0.041
`Meta-Llama-3-70B-Instruct` ICL	0.263±0.118	⇑ 0.347±0.063	0.129±0.118	↑ 0.269±0.234
`Meta-Llama-3-70B-Instruct` CoT	0.242±0.121	↑ 0.383±0.047	0.243±0.096	↑ 0.381±0.135
`gpt-3.5-turbo` ICL	0.319±0.057	↑ 0.357±0.066	0.167±0.236	⇑ 0.468±0.100
`gpt-3.5-turbo` CoT	0.253±0.068	⇑ 0.361±0.057	0.260±0.131	↑ 0.350±0.048
`gpt-4o-mini` ICL	0.228±0.188	⇑ 0.399±0.014	0.394±0.193	↓ 0.245±0.204
`gpt-4o-mini` CoT	0.296±0.039	↓ 0.287±0.053	0.166±0.060	↑ 0.222±0.057

Table 1: Comparison of prior and posterior F1 score of authority and equality for annotator `01` of MFRC using ICL or CoT. ↑ / ↓ : increase/decrease with overlapping ranges, ⇑ / ⇓ : increase/decrease with non-overlapping ranges.

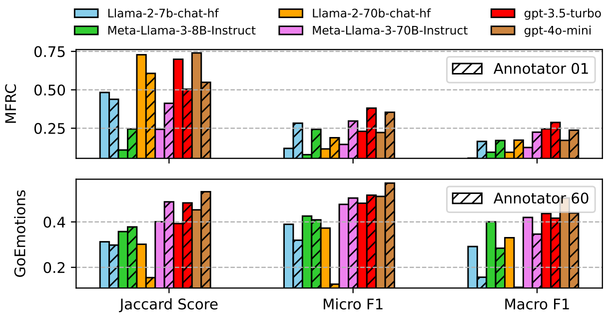

Figure 7: Performance comparison across LLMs using ICL when using aggregated labels and using best-case scenario individual annotator labels.

4.6 Best Performing Annotators

Finally, and as a best case scenario, we present the performance of the best annotator with ICL (based on performance across models as can be seen in Figures 3 and 4) and compare that to the aggregate. Results are shown in Figure 7. The benefits from modeling a specific annotator are evident, as we observe large gains in performance in two out of the three metrics, a finding that is consistent across all models except for LLaMA-2 70b. This further emphasizes that modeling individual perspectives instead of aggregates is beneficial.

5 Conclusion

In summary, we see that aggregated labels tend to align less with the LLMs' prior for each task. Furthermore, and in spite of worse aligned annotators receiving larger posterior performance increases, the aggregate posterior appears to collapse to the prior, resulting in significantly worse performance to several annotators. This result indicates that the majority is not necessarily well-aligned with models. Then, it is evident that interpretable and consistent mappings can be modeled by LLMs and improve upon the prior, even though the model might not align with the specific annotator a priori. Finally, we see that individuals do indeed result in better performance on the task.

Given the commonsense reasoning capabilities of LLMs, the emotional and moral capabilities of LLMs demonstrated in settings different from traditional machine learning settings (Tak and Gratch, 2024), as well as best-case results, we conclude that the **aggregation process introduces artifacts** in the labels that cause LLMs to ignore the mapping as noise. It is interesting to note that in our prompts, the aggregate in rarely a hodgepodge of disparate opinions, wherein the aggregate does not match any individual, but factors like different annotator mixtures, especially between train and test splits, as well as different annotator groups prevailing in different examples, introduce sufficient noise in the annotations that cannot be modeled with ICL and CoT. That being said, the performance gap with gradient-based methods remains large, suggesting that other factors like task complexity also majorly account for the observed biasing effects.

Finally, we question what it means to model these aggregate opinions. Namely, since they might not reflect the opinion of any individual, even if in a minority of cases, then what rationale should be provided and how should it be generated in a sound manner? We advocate, therefore, as previous work has done (Prabhakaran et al., 2021; Dutta et al., 2023), for releasing and modeling annotator-level labels instead of aggregates, and suggest that the field of **subjective modeling should move away from aggregate modeling** in the age of LLMs and of more elaborate modeling methods such as CoT.

6 Limitations

Given our constraints to standardize the prompt and remove other degrees of freedom that can constitute potential confounding factors in our evaluations, but also for computational efficiency, the evaluations sets contain a small number of examples (namely, 100 and 71 for the triplets in GoEmotions, and 100 for both groups in MFRC), increasing the noise in our findings. Nonetheless, this practice has become standard in the evaluation of LLMs.

A potential confounding factor that we do not control for is the quantization, as previous work has reported significant decreases in performance from it (Marchisio et al., 2024). We note, first, that there is no reason for the quantization to affect our results in a nonuniform way, second, that we perform the quantization because of obvious computational constraints, and third, it is possible that even some API-based models are served quantized (e.g., Turbo 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 performance of LLMs in these subjective tasks compared to other models, but rather to compare within the LLMs themselves. Nonetheless, we perform experiments with various levels of quantization of LLaMA-3 8B, and show the results in Table 3 in the Appendix.

In our study of specific labels and the effects of what we perceive as consistency on the performance of LLMs (Section 4.5), a potential confounding factor in analysis is the increased frequency of the label, since the studied annotator was more sensitive to specific stimuli in the input, as described. We specifically chose to perform a more detailed analysis in this positive pattern, however, because we expect the models, as did we, find it easier to distinguish positive patterns in the data.

It is important to note that we have performed experiments in only two problems and benchmarks, as we opted in favor of presenting more LLMs. Therefore, our findings may not generalize to other, highly subjective tasks. Furthermore, other datasets with stricter annotation manuals that aim to resolve all ambiguities may not present similar behavior, as annotator agreement is artificially raised by removing some of the subjectivity of the semantic interpretations of the annotators in favor of following stricter, highly specific instructions.

We also want to note that we do not perform Bonferonni correction across models and datasets given the small number of datapoints we have to compute our correlations, yet we believe it is important to highlight the settings with smaller p-values.

Moreover, datasets with data derived from social media have been criticized as lacking the context for a model—or even humans—to make appropriate judgments about the emotion or morality expressed in them (Yang et al., 2023), and techniques to evaluate the correctness of the labels in a dataset have been designed to discard noisy samples (Swayamdipta et al., 2020; Mokhberian et al., 2022). Given the subjective nature of the task and the lack of context, such tools could be used to perhaps improve performance. That being said, the interpretation by humans is sufficiently consistent for Demux to achieve better performance than every LLM. We also note that by removing ambiguous examples, we may also remove that which makes these tasks challenging (Aroyo and Welty, 2015).

Finally, while we carefully control the experimental setting to control for confounding factors, we do not actually perform causal interventions, and consequently present correlational evidence.

7 Ethical Considerations

Our focus on traditional machine learning benchmarks, as well as our takeaways, should complement and not compete with quantifying bias in LLMs using other tools and techniques (Caliskan et al., 2017; Gonen and Goldberg, 2019; Ferrara, 2023; Abdurahman et al., 2024). It is also important to emphasize that improving affective and moral capabilities in LLMs entails perils, since better catering the emotional and moral responses to more contexts and personalizing them to specific individuals can lead to improved manipulation of users by LLMs.

Acknowledgements

This project was funded in part by DARPA under contract HR001121C0168, and in part by NSF CIVIC.

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. 3(7):pgae245.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. 36(1):15–24.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. 33:1877–1901.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Stephanie CY Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X Wang, Aaditya K Singh, Pierre Harvey Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*.

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, Niyantha Maruthu Pandiyan, Kristina Lerman, and Shrikanth Narayanan. 2025. Larger language models don't care how you think: Why chain-of-thought prompting fails in subjective tasks. In *ICASSP 2025-2025 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 *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. 114(38):E7900–E7909.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. 10:92–110.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. 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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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. *Preprint*, arXiv:2311.00203.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. 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*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

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.

Shivanshu Gupta, Clemens Rosenbaum, and Ethan R Elenberg. 2023. GistScore: Learning better representations for in-context example selection with gist bottlenecks.

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.

Jannik Kossen, Tom Rainforth, and Yarin Gal. 2023. In-context learning in large language models learns label relationships but is not conventional learning.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback.

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

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning" learns" in-context: Disentangling task recognition and task learning.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. 1(8):9.

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*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

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.

Ala N Tak and Jonathan Gratch. 2024. Gpt-4 emulates average-human emotional cognition from a third-person perspective. *arXiv preprint arXiv:2408.13718*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and Shruti Bhosale. 2023. Llama 2: Open foundation and fine-tuned chat models.

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.

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.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently.

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.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. 36.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*.

A Annotators

We reiterate that we group annotators together based on their overlap in the train and evaluation sets so as to standardize the prompt across them. We use 5 annotators in MFRC, one triplet and one pair, both with enough common examples in the evaluation set to present somewhat robust results. For both annotator groups, we use 100 evaluation examples. For GoEmotions, we use 6 annotators, as these were the only groups with enough evaluation examples, 100 and 71 respectively. We did not find any possible quadruplets or larger groups.

B Annotation Details

Two research assistants trained on emotions and moral foundations using the annotation manuals provided by the dataset authors and with common instructions independently generated the reasonings for the annotations in the datasets. We then check for the consistency of the reasoning chains between the two annotators qualitatively (manual checks) and quantitatively (checking the consistency of the predictions when using either annotator's reasonings). Note that we do not use model-generated CoT because we noticed from our experiments that LLM explanations are shallow and can easily miss more complex emotional and moral expressions.

C Toy Example for Aggregation Artifact

In this section, we present an toy example to elucidate the notion of inconsistencies that can be created from aggregation, presented in the main text. Consider an example dataset where the US political affiliation is the major factor dictating the labels provided by annotators. Due to random assignment of examples to annotators, some examples may be assigned more annotators leaning Democrat, and similar for Republican. Therefore, the aggregated dataset will contain some annotations reflecting Republican views, and some Democrat views, creating inconsistencies in the training data, or worse,

Prompt examples

ICL

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

Input: Did anyone watch Nigel Farage rebuke the EU? It was quite interesting!
Moral Foundation(s): none

Input: Because Le Pen is alt right and is very dangerous for the peace of europe and the world.
Moral Foundation(s):

CoT

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

Input: Did anyone watch Nigel Farage rebuke the EU? It was quite interesting!

Reasoning: The author expresses interest in Farage's actions but not any moral sentiment towards it.

Moral Foundation(s): none

Input: Because Le Pen is alt right and is very dangerous for the peace of europe and the world.

Reasoning:

Table 2: Examples showcasing the prompt format in In-context Learning and Chain-of-Thought prompting.

between training and test data. Critically, trying to model political affiliation, the major factor of this setting, becomes impossible from this aggregated data, hence the "inconsistency" of the annotations.

D Prompt

For completeness, we showcase example prompts we use in MFRC for ICL and CoT that illustrate the prompt format we utilize across all experiments. One example for each setting is shown in Table 2. We note that the prompts do not utilize the conversational format even though it is available across all our models because we found it to work worse in terms of performance (in terms of the performance of the aggregate in GoEmotions) and ability to decode (e.g., including explanations without being prompted to, diverging in terms of output format, or even predicting emojis rather than emotion words) compared to the used one.

Q.	**Micro F1**	**Macro F1**	**JS**
Aggregate			
4 bit	$0.078_{\pm 0.026}$	$0.094_{\pm 0.034}$	$0.108_{\pm 0.013}$
8 bit	$0.124_{\pm 0.036}$	$0.137_{\pm 0.017}$	$0.241_{\pm 0.071}$
16 bit	$0.131_{\pm 0.047}$	$0.150_{\pm 0.014}$	$0.219_{\pm 0.061}$
Annotator #1			
4 bit	$0.090_{\pm 0.030}$	$0.111_{\pm 0.033}$	$0.114_{\pm 0.046}$
8 bit	$0.123_{\pm 0.061}$	$0.155_{\pm 0.049}$	$0.240_{\pm 0.108}$
Annotator #2			
4 bit	$0.243_{\pm 0.038}$	$0.170_{\pm 0.024}$	$0.245_{\pm 0.038}$
8 bit	$0.208_{\pm 0.059}$	$0.228_{\pm 0.022}$	$0.312_{\pm 0.042}$
Annotator #3			
4 bit	$0.122_{\pm 0.043}$	$0.122_{\pm 0.036}$	$0.154_{\pm 0.090}$
8 bit	$0.126_{\pm 0.018}$	$0.135_{\pm 0.020}$	$0.247_{\pm 0.072}$

Table 3: Performance of LLaMA-3 8B with various levels of quantization (Q.) on MFRC.

E Quantization

We perform experiments with LLaMA-3 8B with less quantization to gauge at the effects that it has on our results. We present our findings in Table 3. As expected, performance slightly improves with less quantization for our metrics of interest, yet the comparative phenomena we study still seem to hold from this small-scale analysis.

F Full Results

In Figures 8, 9, 10, 11, and 12, we see the equivalent results to Figures 2, 3, 4, 5, and 6 respectively, but each annotator is assigned their own marker. Therefore, these more detailed figures allow us to track the performance of different annotators across models and settings.

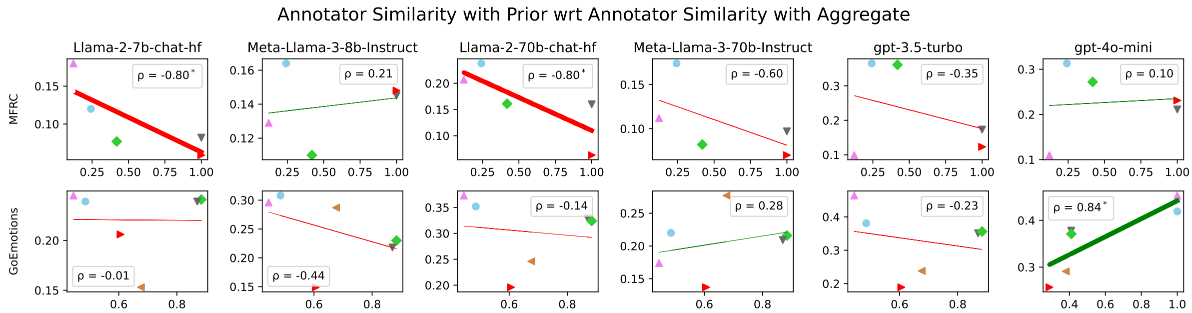

Figure 8: Scatter plot of annotator similarity with aggregate and with prior, with correlation and line fit of data shown. * and bold lines: $p < 0.05$.

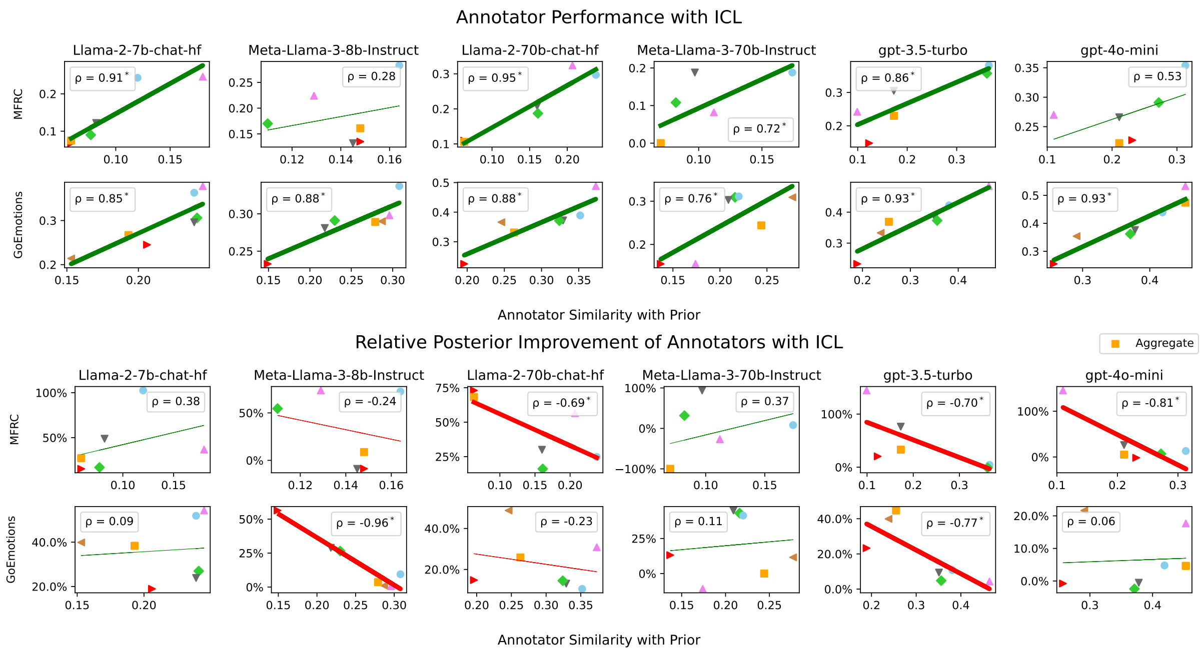

Figure 9: **In-Context Learning** performance for annotators and aggregate and their relative improvement compared to the model's prior as a function of the similarity of each with the model's prior. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

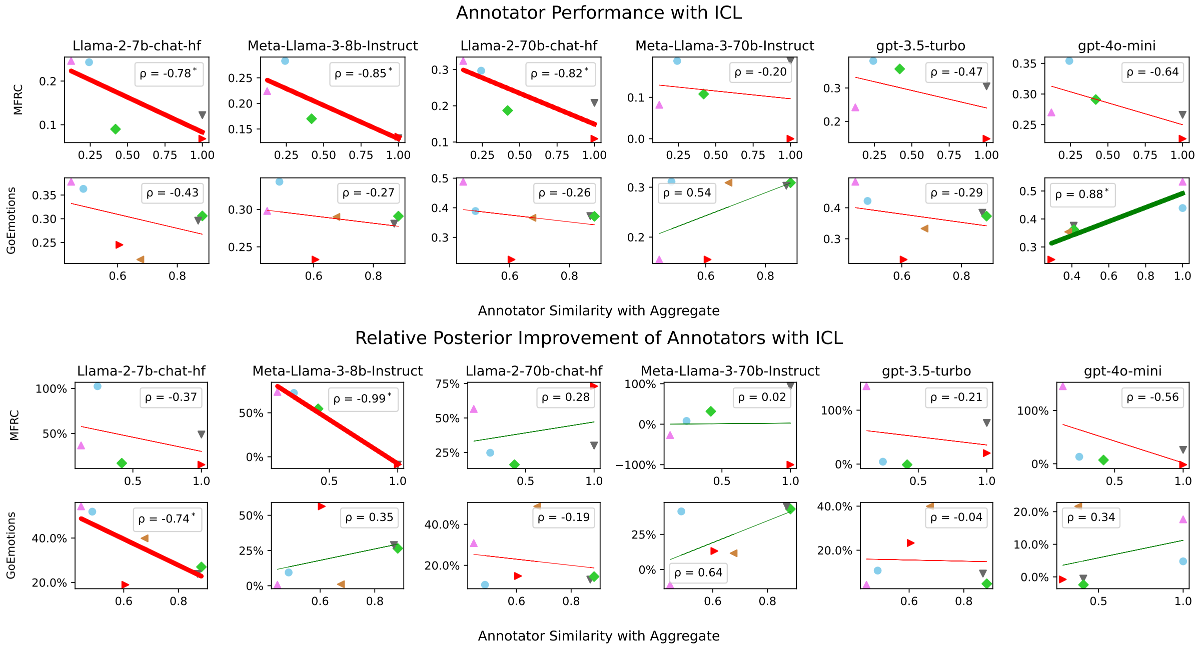

Figure 10: **In-Context Learning** performance for annotators and their relative improvement compared to the model's prior as a function of the similarity of each annotator with the aggregate. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

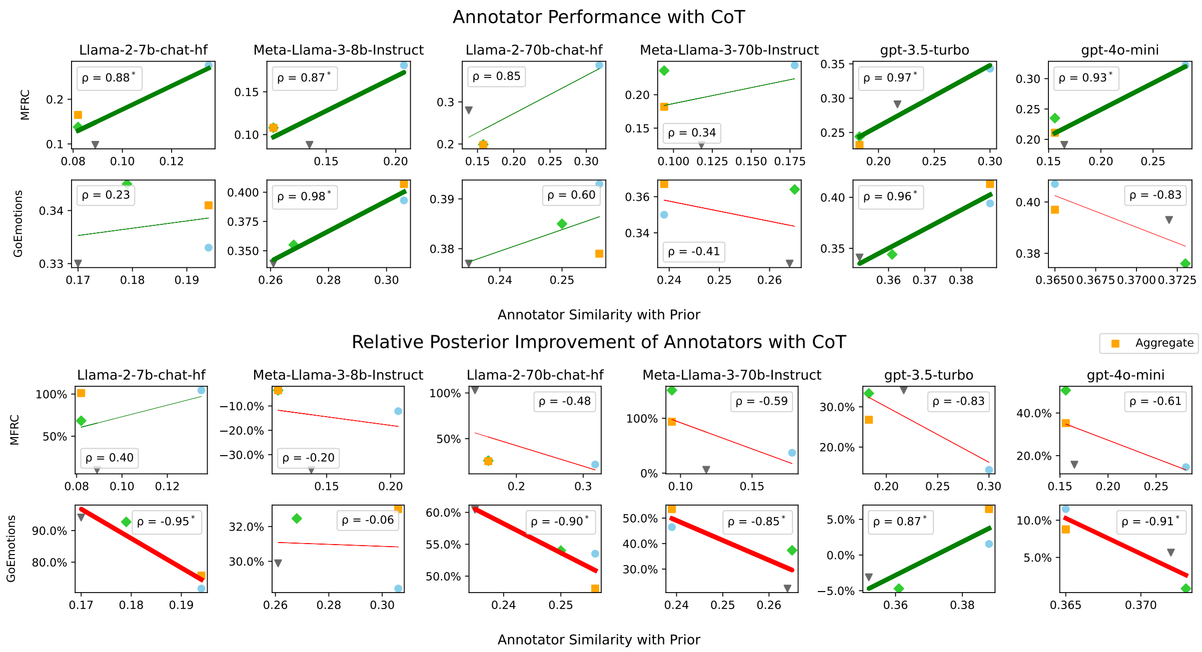

Figure 11: **Chain-of-Thought** performance for annotators and aggregate and their relative improvement compared to the model's prior as a function of the similarity of each with the model's prior. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.

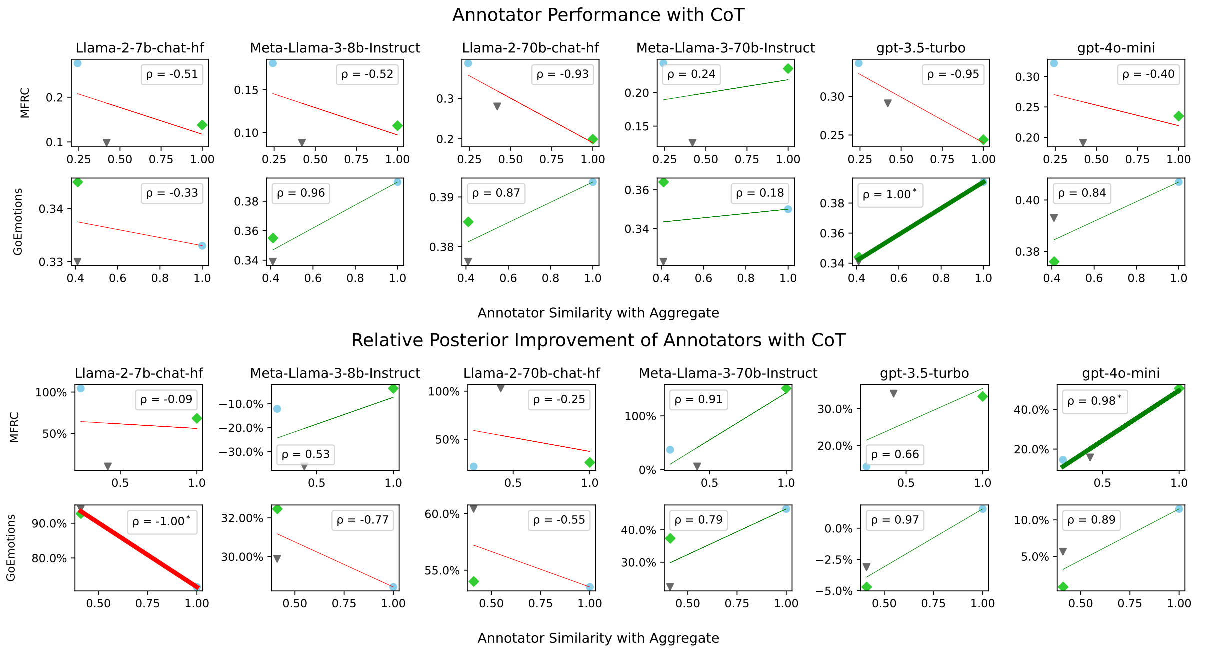

Figure 12: **Chain-of-Thought** performance for annotators and their relative improvement compared to the model's prior as a function of the similarity of each annotator with the aggregate. Correlation and line fit of data also shown. * and bold lines: $p < 0.05$.