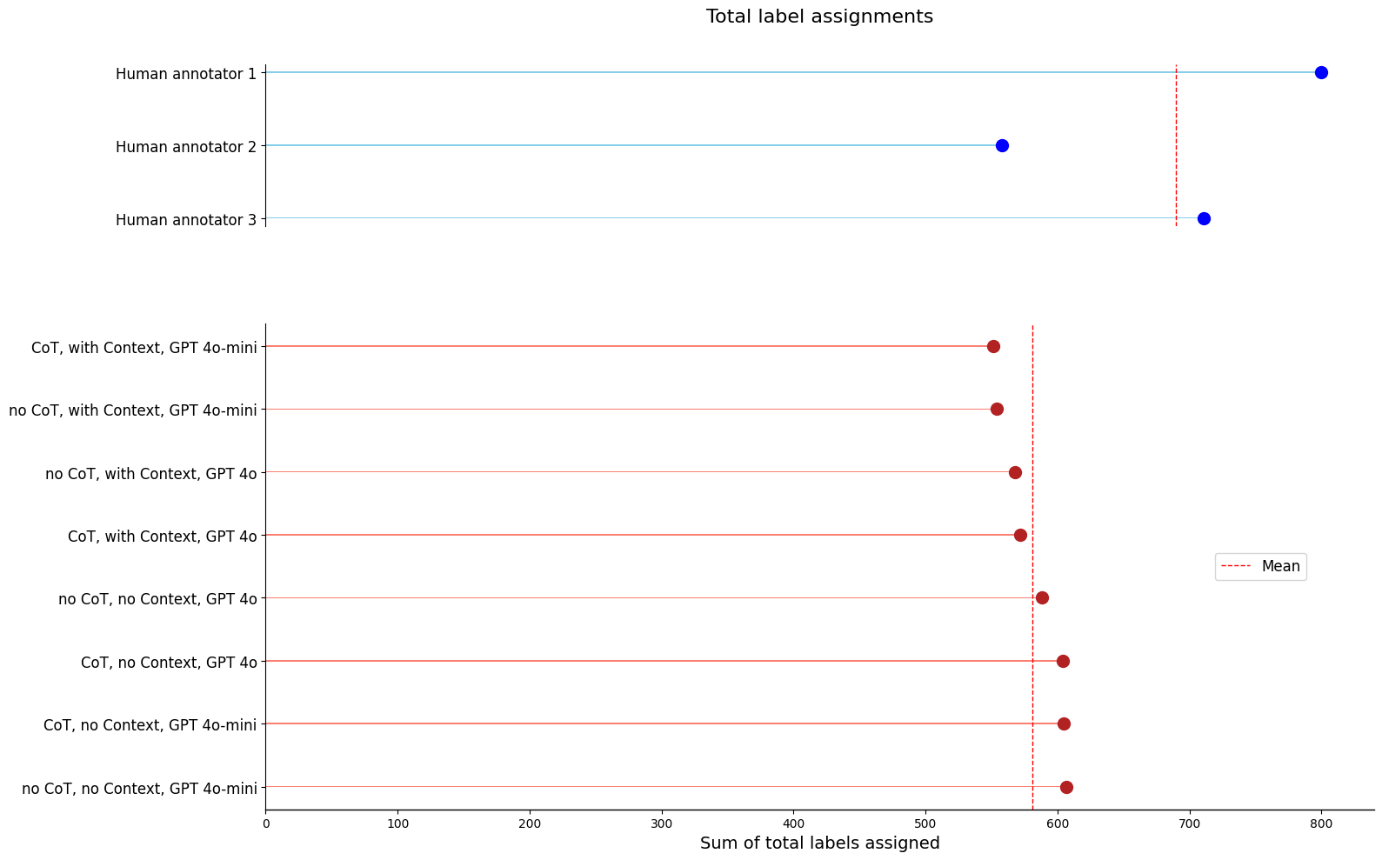
## Supplement 4: Annotation comparisons and algorithm selection

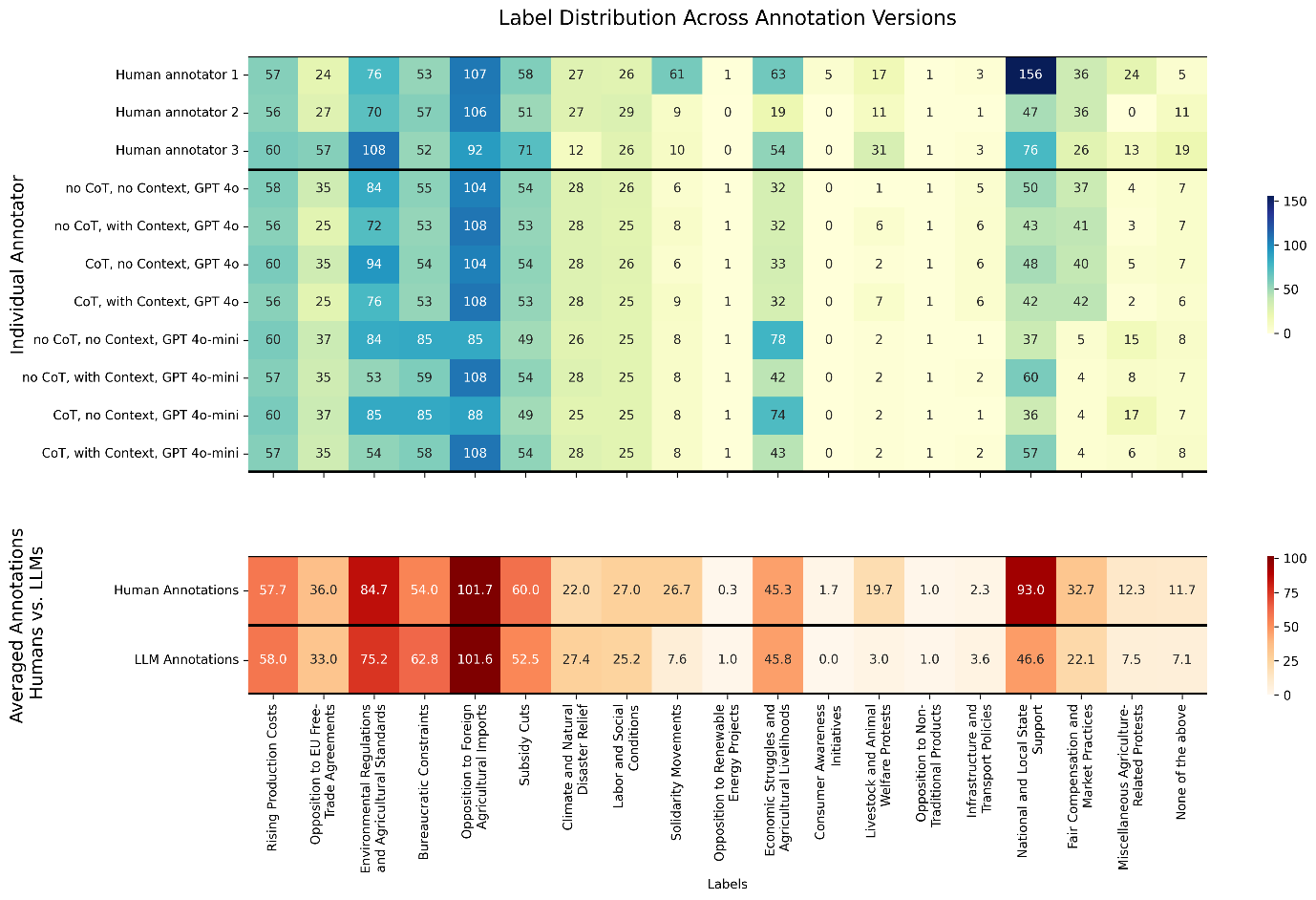
As outlined in the methods section, a major difficulty in finding an appropriate classification algorithm is the absence of a ground truth baseline. Therefore, we set the classification results by three trained human annotators as baseline, against which we compared the LLM classifiers and finally selected our preferred model.

Figures S1 and S2 compare the GPT based annotations with human annotations. We can see that the human annotators assigned considerably more labels than GPT based, especially annotators 1 and 3. Also, those GPT annotators that were not provided with class descriptions assigned in total more labels than those that were.

In terms of assigning labels to each class, we can see many similarities between the different models, e.g. for *Rising Production Costs* and *Opposition to Foreign Imports.* But there are also rather large differences between GPT and human annotators on average, e.g. for *National and Local State Support, Solidarity Movements* (driven by annotator 1). Differences within GPT models, especially with respect to *Economic Struggles and Livelihoods*. We also found that GPT 4o models were relatively homogeneous in terms of individual label assignment counts.

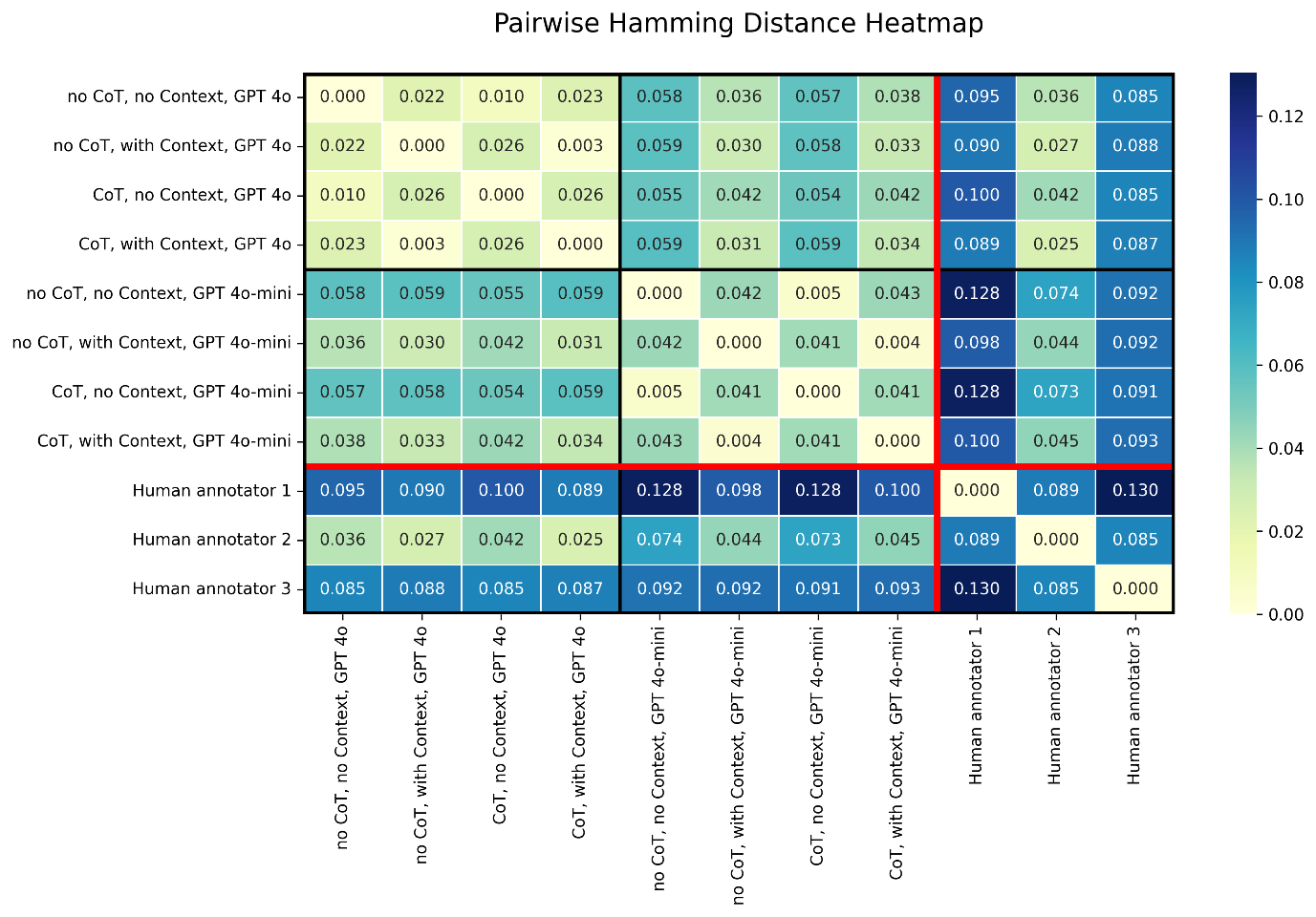


*Figure S 1: Total label assignments on validation subsample.*



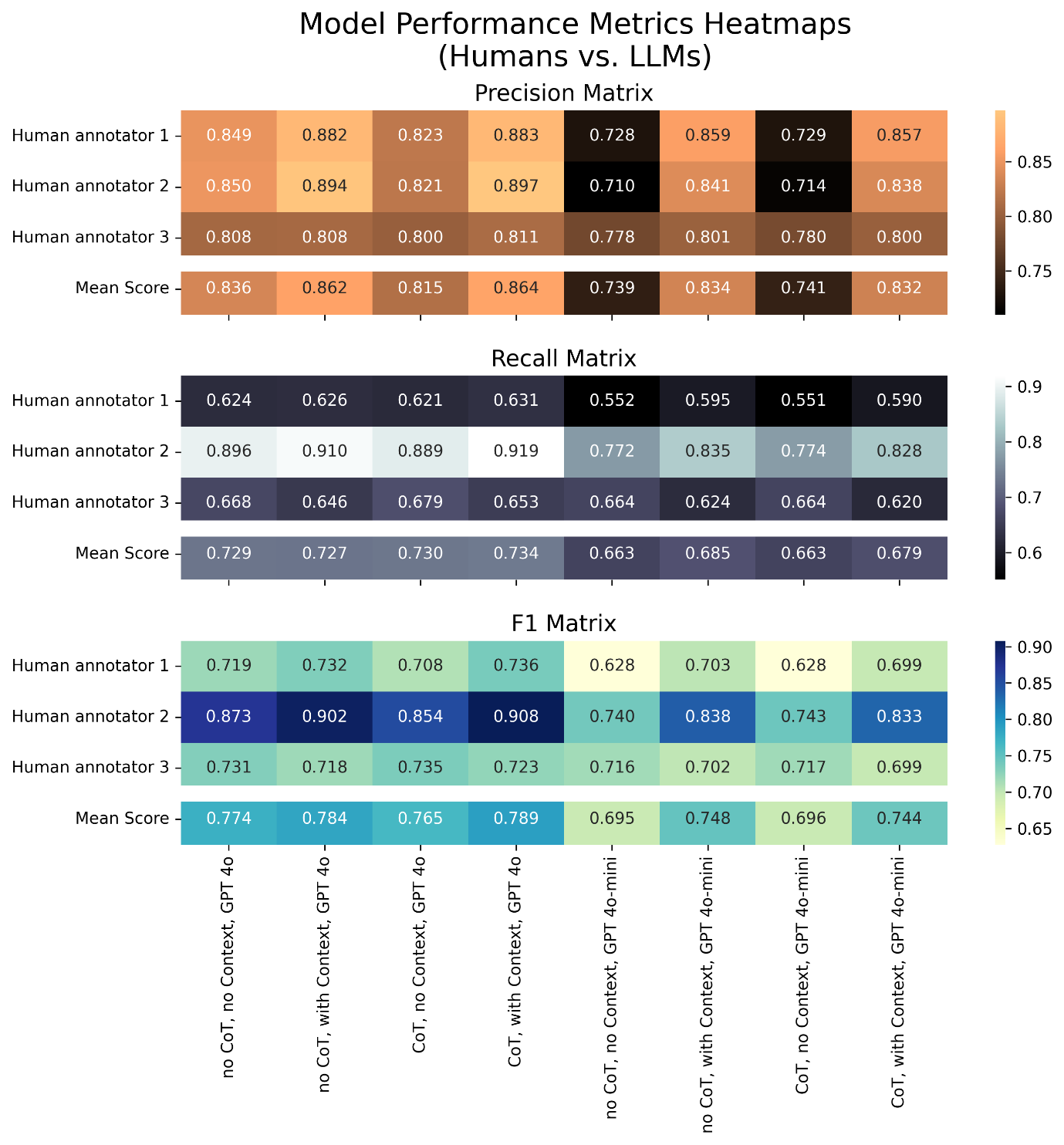
*Figure S 2: Number of annotations per category of reasons.*

To select our preferred AI classifier, we measure inter-classifier agreement measured by pairwise Hamming distances (fig S3). A few observations can be made. The large agreement among GPT 4o annotators from above could be confirmed. Overall, GPT models showed high classification agreement, whereas human annotators had a lower inter-annotator agreement. We found overall a good agreement between GPT and human annotations (comparable to inter-human agreement), especially annotator 2 had a very high agreement with GPT models. On average, the Hamming distance between GPT 4o with provision of text descriptions and chain of thought prompting (CoT, with Context, GPT 4o) and human annotations was lowest, which pointed toward that annotation model (CoT, with Context, GPT 4o) as preferred model to perform the annotation task for all instances.



*Figure S 3: Inter-classifier agreement measured by Hamming distance.*

To further evaluate this, we calculated precision, recall and F1 scores of the GPT models against the human baseline annotations. We found high precision scores, lower but still pretty high recall scores (which is not surprising given that humans assigned considerably more labels overall), and consequently high F1 scores. Furthermore, the result from the Hamming distance comparison could be confirmed. GPT 4o with provision of text descriptions and chain of thought prompting (CoT, with Context, GPT 4o) was the most human-like LLM annotation model.



*Figure S 4: Evaluation scores. GPT vs. human baselines.*

Finally, to obtain a detailed picture of overlaps and divergences between human annotations and annotations for each protest reason by our preferred LLM specification, Supplementary information SI 5 provides reason-specific confusion matrices.