SALOMO: An Annotation Tool For Complex Annotation Tasks With A Large Number of Labels

Tim Menzner **Jochen L. Leidner**
Center for Responsible Artificial Intelligence (CRAI), Coburg University of Applied Sciences,
Friedrich-Streib-Str. 2, 96450 Coburg, Germany
Contact: tim.menzner@hs-coburg.de / leidner@acm.org

Abstract

Manual annotation of linguistic units such as sentences with labels drawn from a large inventory or taxonomy imposes an enormous cognitive load on human subjects. For our demonstration task, we devised a taxonomy of media bias with 37 categories. Selecting the appropriate category (or none) for thousands of news sentences is likely to be tiring and error-prone for humans. To address these type of annotation tasks involving large numbers of labels, we present *SALOMO* (Smart Annotation Labeling On Model Outputs), an annotation tool that pre-selects labels by letting a committee of LLMs make decisions. Human annotators are then tasked mainly with resolving cases where the LLMs disagree. While our tool is independent of any particular task, we describe its design, present a short corpus annotated with a novel fine-grained taxonomy of news bias types as a concrete case study, and demonstrate experimentally both the significant time savings and workload reduction achieved with the pre-selection mechanism, as well as the strong bias it introduces toward the displayed selection. We also provide the mini-dataset of biased sentences and their associated bias types from our experiment.

Keywords: annotation tools, complex coding tasks, news bias annotation, media bias, propaganda, annotated news corpora, LLM for quality management, news bias detection, propaganda identification, natural language processing, news mark-up, data coding, corpus annotation, linguistic resources

1. Introduction

Annotating data is important but difficult work. Without countless humans willing to invest the time and effort to examine vast amounts of data, item by item, and apply their cognitive skills to categorize it, the current “age of machine learning" would not have been possible in the first place. High-quality annotations are the foundation of good quality in machine learning. Therefore, it is of utmost importance to ensure that the annotation process runs as smoothly as possible. This can already be challenging even in seemingly simple cases, where annotators choose from a small set of clear-cut categories and must decide whether each one applies or not. Empirical research has shown that even under such conditions, annotation is often prone to errors, inconsistencies, and considerable variation across annotators (Bernhardt et al., 2022). These difficulties become even more pronounced as the number of categories increases or when categories are ambiguous, making efficient workflows and supportive tools essential.

Now imagine an annotation task where ambiguity plays a central role, for example, due to the nuances and subjectivity of human language, which often make the ideal of unambiguously defined categories unrealistic. Under such circumstances, the time required to carefully consider each category can be expected to increase significantly (Deng et al., 2014). The challenge grows even further when the task involves a large number of categories and when a single item may require assignment to multiple labels rather than just one.

As Miller (1956) showed, human short-term memory is not designed to hold more than about seven chunks of information at once; later research suggested an even lower limit of four (Cowan, 2001).

This is precisely the situation we faced in our annotation work: our goal is to construct a dataset of sentences drawn from online news articles, each of which must be annotated as either unbiased or biased according to one or more categories from our taxonomy, which currently comprises 37[1] distinct types of media bias.(Menzner and Leidner, 2026)

A first pilot experiment, in which a small group of students annotated 25 sentences using an earlier version of our taxonomy (with 31 bias types), revealed the limitations of a traditional annotation approach. On average, participants required about three minutes per sentence, with times ranging from 90 seconds to seven minutes, and they generally described the task as very difficult. This suggests a high likelihood of errors, indeed, inter-annotator agreement was relatively low, though this may also be attributed, at least in part, to the inherent ambiguity of the task or insufficiently detailed guidelines. Carefully considering all 31 categories for each sentence proved to be extremely demanding.

These findings indicated that the task was not scalable, not only because of its inherent complexity but also because longer annotation times can

[1] 38 in the final version

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6312–6321
11–16 May 2026.

directly translate into higher costs. While some studies did not find such effect (Mason and Watts, 2010), others established that annotation quality often increases with financial incentive (Laux et al., 2025). Better pay can mean better motivation to carefully think through more complex annotation decisions, so in order to get reliable results, especially for complex tasks, it is safer to pay annotators fairly (besides the obvious ethical reason, of course). However, if a high task complexity means each individual sentence takes too long, the required time per sentence would place an unsustainable burden on resources.

Therefore, we decided to pursue a different approach: harnessing the growing capabilities of large language models (LLMs) to handle complex linguistic tasks, while still keeping human annotators in the loop for quality control. In our setup, a committee of LLMs first makes a pre-selection of categories, which then serves as the basis for human evaluation. To further speed up and streamline this workflow, we developed *SALOMO*, an annotation tool specifically designed for scenarios with a large number of categories and tailored to facilitate work with such pre-selected options.

In the following, we present *SALOMO* in more detail, outline its key features, and describe how it supports efficient annotation with pre-selected categories. The tool will be made available via GitHub to encourage transparency, reproducibility, and community adoption. We will also describe our general process of LLM based pre-selection, serving as foundation of this tool.

2. Related Work

2.1. LLMs for Annotation

The use of large language models (LLMs), particularly generative pretrained transformers, for data annotation, either autonomously or with a human in the loop, is not a new idea. To our knowledge, the earliest studies exploring this approach appeared shortly after the release of GPT-2 and GPT-3. For example, Wang et al. (2021) demonstrated that GPT-3 could achieve annotation quality comparable to humans across six different natural language tasks, and that combining model-generated annotations with human-labeled data yielded even greater improvements.

A comprehensive overview of research in this field, from its early beginnings up to 2024, is provided in two surveys by Tan et al. (2024) and Pavlovic and Poesio (2024). As might be expected given the varying circumstances, such as differences in LLMs, task complexity, and prompting strategies, the findings vary. While most studies report that LLM-generated annotations can serve as a reliable substitute for human ones, sometimes even outperforming them (Törnberg, to appear), others arrive at more mixed conclusions (Belay et al., 2025). The overall consensus appears to be positive, however, indicating that LLMs can indeed be useful for annotation work in many situations. Their usefulness in settings where several labels from a large pool can be assigned was demonstrated by Niu et al. (2024), who had LLMs classify sentences according to the 27 categories from the GoEmotion dataset, while Ma et al. (2025) focused on the inner workings of LLMs when assigning multiple labels in contrast to single-label assignments.

The performance of combining LLM annotation with a human in the loop is to a large part connected to the performance that the LLM would achieve on its own. If the LLM is already very good, the overall quality of the annotation can benefit, if not, having an LLM in the loop might be worse than a human alone (Lippolis et al., 2025).

This is probably connected to the fact that human annotators have shown a bias to just go with the LLM suggestions for the most part (Schroeder et al., 2025), so if these suggestions are good, human feedback might be the icing on the cake, while if they are bad, the whole annotation is built on sand.

Findings about the workload and time savings remained mixed, while Schroeder et al. (2025) did not find any significant reduced in annotation times, other studies like (Wang et al., 2024) found those effects.

Interestingly, some guidelines for designing human-based annotation approaches may not be necessary when creating annotation tasks that rely solely on LLMs, while others still apply, or may even be unique to LLMs. For example, when items are shown independently, there is no need to account for order effects on annotator behavior, as is the case with humans (Beck et al., 2024). However, research has also shown that the order of instructions and labels within prompts can significantly influence LLM labeling decisions (Guan et al., 2025).

2.2. Annotation Tools

The ideas of hierarchical annotation, and in pursuing it, to reduce annotator workload, has long been one of the core ideas behind the design of annotation tools (Henley and Piorkowski, 2024).

Leidner and Jung (2024) use a hierarchical tree structure that capturing multi-level (geographical) entity relationships, and this could be used to improve the accuracy and efficiency of geographic text annotation with spatial footprints (“toponym resolution") via context-aware disambiguation. Also, LLM assistance has already been incorporated into several popular annotation tools, including INCEpTION (Klie et al., 2018), Label Studio (HumanSignal, 2025), and Label Sleuth (Shnarch et al., 2022).

The nature of these integrations varies widely, from providing LLM-based annotation suggestions to automatically training new classifiers in the background, demonstrating many different ways LLMs can enhance the annotation workflow.

Perhaps the closest to our work is Kim et al.'s MEGAnno+ (Kim et al., 2024), which is based on a Jupyter notebook and allows human review of model outputs with an in-notebook verification interface.

Our work differs in that we fully embrace a "human-as-tiebreaker" approach: while human intervention is still possible if an annotator notices an obvious error that should be corrected, the default mode is to treat overlapping decisions from two different LLMs as already confirmed. Human annotators are only asked to resolve labels where the LLMs disagree. We also incorporated design improvements based on feedback from our manual annotation experiment described in Section 1, such as displaying guidelines when hovering over a category. Additionally, recognizing that human annotators, and even researchers without a technical background who may need to annotate (or "code", as it is often called in media research, for example), might not be highly tech-savvy, we prioritized ease of use. The tool is deployed as a single HTML5 page that can be run locally in every modern browser, minimizing technical barriers.

3. The *SALOMO* Annotation Tool

3.1. LLM Pre-annotation

Since the goal is to simulate the diversity that would naturally arise from employing different human annotators, each with their own subjective annotation style, it makes sense to use a pair of LLMs that differ sufficiently in their annotation behavior. At the same time, they should demonstrate a strong enough understanding of the task to produce reliable results.

For this experiment, we employed OpenAI Inc.'s GPT-5 and GPT-o3. Both models offer advanced reasoning capabilities, which are especially valuable for a complex task such as bias annotation, yet they exhibit distinct annotation tendencies. Specifically, GPT-o3 tends to be more conservative in label assignment: In our experiment 4, it assigned about 1.68 labels on average per sentence, whereas GPT-5 showed a tendency to assign more labels, averaging around 3.0.

We also tested smaller models, including open-source alternatives, but observed clear performance gaps compared to the larger models. Consequently, we decided against their use in this setup.

In terms of inter-annotator agreement between the two models, differences in their tendencies regarding the number of labels to assign prevent the agreement metrics from reaching very high values. This results in a Jaccard similarity of (0.503) on average. Similarly, using Cohen's κ for multi-label annotations, treating each label as a binary decision, the micro-averaged Cohen's κ across all labels and items was $\kappa = 0.581$. The deflation of traditional inter-annotator agreement metrics in multi-label settings is a well-known problem (Marchal et al., 2022). Therefore, we also employed the bootstrapping mechanism presented in the referenced paper, which confirms the moderate agreement with an adjusted F1 for the bootstrap of 0.514 (500 samples).[2]

At present, LLM-based annotation is carried out externally via a Python script, though integration into the tool itself is considered for future work. Both LLMs receive the same prompt, which provides (i) a general definition of media bias, (ii) a brief description of each category, and (iii) an example for each category. These elements will also be shown to human participants when they perform the annotation task in experiment 4. The models are then instructed to return their annotation in JSON format.

The outputs are aggregated into JSONL files, where each line corresponds to one annotation result. From these outputs, we construct a combined result with two fields: one containing the union (labels selected by any of both models), and another containing the differences (labels selected by only one model).

3.2. SALOMO

These fields can then be loaded into *SALOMO*, where the annotator can work through the text sentence by sentence. The interface displays the LLM pre-selection, highlighting where the models agree and disagree, and allows the annotator to resolve conflicts or add and remove labels as needed. Adding a new label is made simple through autocomplete suggestions for annotator input, while removing a category is as easy as clicking a small "x" next to it.

To help annotators navigate the large number of categories, we integrated tool-tips that appear when hovering over a category in any context. In our setup, these tool-tips provide a short definition and an example sentence for each category. For

[2]We also developed our own metric, the MultiLabel agreement. However, the corresponding paper has not yet been peer-reviewed, so we decided against reporting it officially, as it is of limited use right now without the paper being available. But for future reference, the values would be $A_{ml}^{(k)} = 0.511$, $A_{ml_{\min}}^{(k)} = -0.248$, $A_{ml}^{(ac1)} = 0.945$, $A_{ml_{\min}}^{(ac1)} = 0.860$ with $p_r = 0.063$.

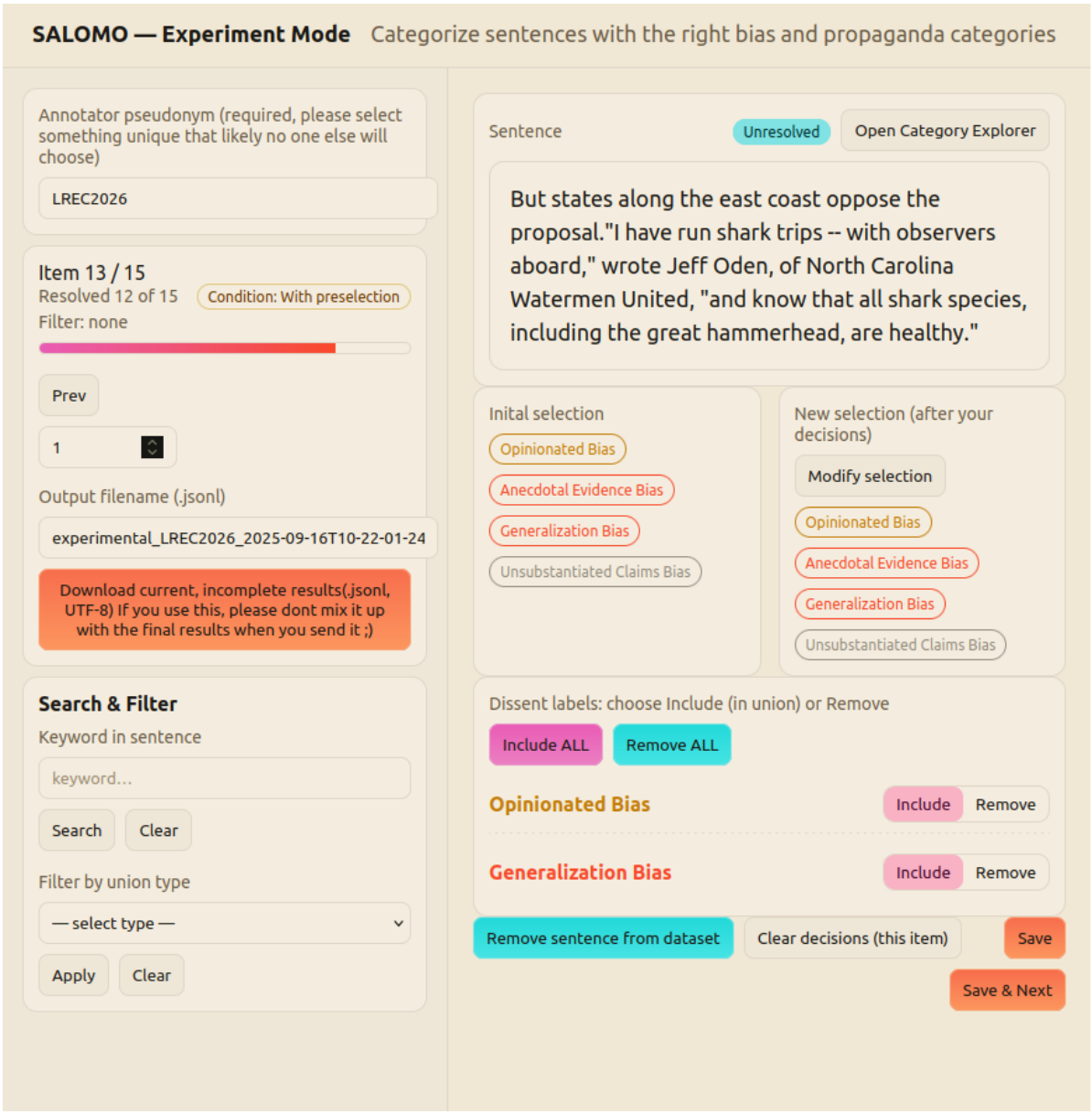

Figure 1: Screen Capture of *SALOMO* in the hard-coded experimental mode, colors inverted from original dark-mode design for better readability in print

better visual orientation, categories are also color-coded according to the higher-level groups they belong to. Both the tooltip content and the color scheme can be adjusted for different annotation tasks via an external configuration file (though they were hard-coded for the experiment described in 4, to limit the number of files that potentially not-so-tech-savvy participants had to handle.)

Besides, the tool-tips, an overview of all categories is also available from all menus via a button.

It is also possible to filter sentences by specific categories or search for a particular sentence using keywords.

A screenshot of *SALOMO* displaying a sentence with pre-selection in the hard-coded version used in the experiment in Section 4 is shown in Figure 3.

Annotators can export their final annotations as JSONL files. These saved results include additional metadata, such as whether a user overrode the model's pre-selection, for example, by adding a category not suggested by either model, or by removing one suggested by both. The specific categories affected by these overrides are also recorded.

For detailed technical instructions on how to use the tool, including the exact field names used in compatible JSONL files, please refer to the project's Git repository.

We chose the name *SALOMO* (short for: Smart Annotation Labeling On Model Outputs), which also refers to the German form of Solomon, the Israelite king renowned in the biblical narrative for his wisdom and ability to resolve conflicts and disputes, as

it alludes to the system's role in mediating between multiple model suggestions.

SALOMO is available to download on git https://github.com/Timperator2/SALOMO

4. Experiment

4.1. Participants

To evaluate whether our LLM-assisted annotation actually reduces time and effort, we conducted a small experiment with n = 21 participants. Of these, 10 identified as male, 8 as female, 2 preferred not to disclose their gender, and 1 identified as non-binary. The age distribution was as follows: 7 participants were between 18 and 25 years old, 5 between 26 and 30, 6 between 31 and 40, 2 between 40 and 50, and 1 participant was older than 50.

Participants also differed in how frequently they consumed news content: 7 reported doing so a few times per month, another 7 a few times per week, 6 almost daily, and 1 only a few times per year. When asked about preferred news sources, 8 participants chose not to provide further details. Among the remaining 13, all named at least one print or online newspaper. The most frequently mentioned outlets were the German left-liberal *Zeit*, the center-left *Spiegel*, and the centrist *Tagesschau* website of Germany's public broadcaster *ARD*, each cited three times. Additionally, 5 participants mentioned using social media platforms such as *Reddit*, *Instagram*, or *TikTok* to stay updated on news.

When asked to rate their English proficiency on a scale from 1 ("very low") to 5 ("very high"), 1 participant selected 2, 9 participants selected 3, another 9 selected 4, and 2 participants rated their level as 5.

As a small token of appreciation, we raffled off three little thank-you gifts among all participants.

4.2. Setup

We randomly selected 30 sentences from our work in progress dataset with biased sentences and split them into two batches of 15 sentences each. One batch (WITH_PRESELECTION) was assigned to display the LLM pre-selection in the tool (participants were not informed that it was LLMs that did the reselection though), while the other batch was presented without pre-selection (NO_PRESELECTION), requiring participants to assign categories from scratch. Which sentence was assigned to which batch was randomized for each participant, as were the order of the batches.

Before the experiment, participants had unlimited time to familiarize themselves with the annotation categories. Prior to each batch, they could practice the workflow using one additional example sentence (not included in the batch) to ensure comfort with the interface. Between the two batches, participants could take a break of any length, although they were encouraged to complete the experiment in a single session aside from this break.

After completing both batches, participants were asked to rate their confidence in their annotations and the perceived difficulty of the task on a Likert scale from 1 to 5 for each condition. We also collected some demographic information, including age bracket and news consumption habits, and provided an opportunity for general feedback.

For reproducibility, a version of *SALOMO* with the hard-coded annotation task used in the experiment, along with the instructions provided to participants, has been made available via Git. The same repository also contains the resulting dataset with all participant annotations, along with the "majority opinion". defined as all types named by more than half of the annotators. A separate majority opinion is provided for each condition, as well as an overall one that includes all types appearing in both condition-specific majority opinions: https://github.com/Timperator2/SALOMO-Experiment.

5. Results

5.1. Quantitative

5.1.1. Completion Time

Average completion time per item was lower in the WITH_PRESELECTION condition ($M = 84.8$ s) than in the NO_PRESELECTION condition ($M = 118.3$ s). The mean paired difference (NO – WITH) was 33.5 s, 95% CI $[-1.1, 68.1]$. Overall, 14 of 21 participants (67%) were faster under the WITH_PRESELECTION condition.

A Shapiro–Wilk test indicated that the differences were not normally distributed ($p < .001$). Consequently, a Wilcoxon signed-rank test was used, revealing a significant difference, $W = 49.0$, $p = .0195$.

Median times in seconds per sentence for all participants and the trend-line between conditions are shown in Figure 2.

5.1.2. Number of Labels per Item

Participants generated more labels per item with pre-selection ($M = 3.35$) than without ($M = 1.99$). The mean paired difference (NO – WITH) was -1.36, 95% CI $[-1.77, -0.95]$.

A Shapiro–Wilk test did not reject normality ($p = .300$); thus, a paired-samples *t*-test was conducted, showing a highly significant difference, $t(20) = -6.94$, $p < .001$. Nineteen of 21 participants (90%) labeled more items under the WITH_PRESELECTION condition.

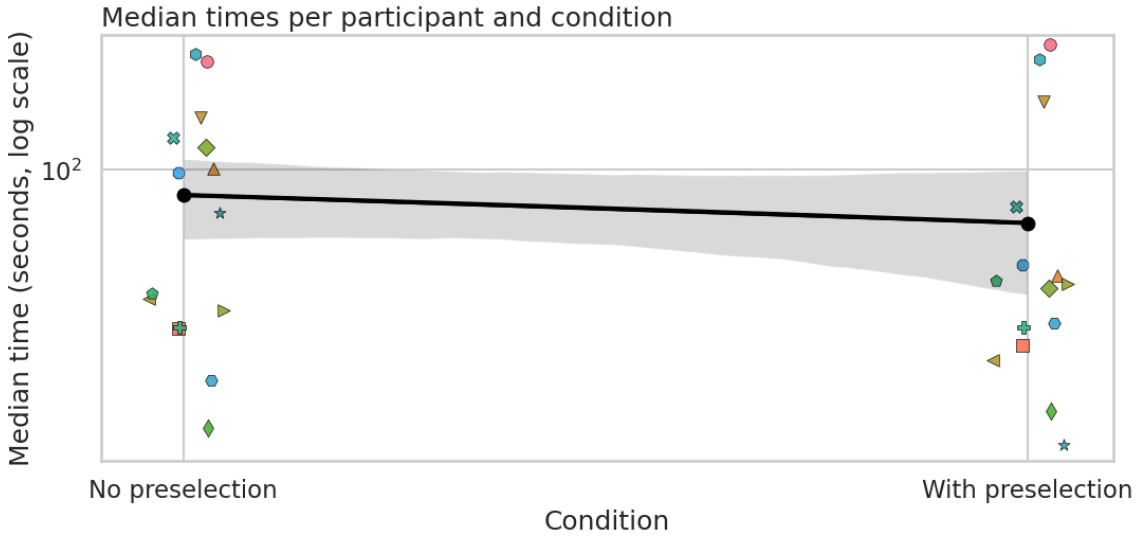

Figure 2: Median times in seconds per sentence for every participant and condition

5.1.3. Perceived Difficulty

Perceived difficulty ratings were lower with pre-selection ($M = 2.52$) than without ($M = 3.86$), indicating that participants found the pre-selection condition easier. The mean paired difference (No – With) was 1.33, 95% CI $[0.73, 1.93]$.

Due to the ordinal nature of the data, a Wilcoxon signed-rank test was applied, showing a significant difference, $W = 10.5$, $p = .0015$. Sixteen participants rated No_Preselection as harder, one found With_Preselection harder, and four rated them equally difficult.

5.1.4. Confidence in Labels

Participants reported greater confidence in their labels under the With_Preselection condition ($M = 3.71$) compared to the No_Preselection condition ($M = 3.05$). The mean paired difference (With – No) was 0.67, 95% CI $[0.05, 1.28]$.

Given the ordinal nature of the data, a Wilcoxon signed-rank test was used, indicating a significant effect, $W = 27.5$, $p = .031$. Eleven participants felt more confident with pre-selection, five without, and five reported equal confidence.

5.1.5. Label Diversity

Participants used more unique labels with pre-selection ($M = 18.95$) than without ($M = 14.86$), a significant difference, $t(20) = -3.43$, $p = .0027$. Thus, pre-selection increased both the number and diversity of labels applied.

Label overlap across conditions was moderate. The mean Jaccard similarity across all sentences was 0.303, and the overall label-set similarity $J = 0.756$, indicating considerable overlap but also some condition-specific labels (eight unique to No_Preselection, two to With_Preselection).

Pre-selection shifted labeling behavior toward stylistic and subjective categories, such as *Word Choice*, *Opinionated*, and *Vagueness Bias*, while reducing more inferential types like *Claim & Blame* or *Source Selection Bias*. Sentence-level comparisons confirmed that pre-selection sometimes radically altered which biases were identified.

Overall, pre-selection broadened individual label diversity while simultaneously promoting greater convergence among annotators. To quantify agreement, we chose Krippendorff's α because sentences were randomly assigned to one of two conditions, resulting in participants effectively missing 50% of sentences per category; Krippendorff's α is robust to such missing data(Krippendorff, 2018). Treating each label as a binary decision, α values indicated moderate to high reliability for With_Preselection ($\alpha_{\text{macro}} = 0.545$, $\alpha_{\text{micro}} = 0.799$), compared to near-chance agreement without pre-selection ($\alpha_{\text{macro}} = 0.049$, $\alpha_{\text{micro}} = 0.098$). The bootstrap-adjusted measures corroborated this pattern, showing substantial agreement for With_Preselection ($F_1 = 0.640$) but minimal consensus for No_Preselection ($F_1 = 0.076$).

Together, these findings suggest that pre-selection both expanded the range of labels available and anchored annotators' judgments around a shared stylistic and subjective framing. The result was a labeling process that was simultaneously more diverse and more internally consistent, a pattern mirrored in the subsequent *Stickiness* analyses, where pres-election likewise enhanced label stability and reuse.

5.1.6. Stickiness and Adherence to Pre-selected Labels

(With_Preselection only)

Stickiness (adherence to pre-selected labels) was operationalized as the proportion of final labels that were pre-selected, mathematically equivalent to:

$$\text{precision} = \frac{|\text{final} \cap \text{pre-selection}|}{|\text{final}|}.$$

Additional similarity measures were also computed per annotated sentence.

Participants retained the vast majority of suggested labels (mean precision $= 0.893$) and rarely deleted them (mean deletions $= 0.467$ per item). The Jaccard similarity ($M = 0.774$) further indicates high overall overlap between suggested and final label sets. These findings confirm strong reliance on pre-selection, consistent with the notion of *stickiness*.

5.1.7. Demographic Correlations and Self-Reported Learning

While not directly related to the potential benefits of our tool, several findings from the statistical analysis of demographic correlations offer noteworthy insights. These results may inform practical considerations for similar tasks and annotator selection.

A Spearman rank-order correlation examined associations between demographic variables and task performance. Age was significantly positively correlated with average task time, $\rho = .48$, $p = .029$, indicating that older participants tended to take longer to complete the tasks. English proficiency showed a moderate, non-significant negative association with task time, $\rho = -.42$, $p = .058$, suggesting a trend in which higher English proficiency corresponded to faster task completion.

Because our broader goal concerns the potential of interventions to enhance media literacy, we also asked participants for a subjective self-assessment of whether the forced engagement with the topic had any effect on their awareness or understanding. Although such self-assessments cannot substitute for objective measures, they provide indicative insight.

Participants generally agreed that they learned something new about bias and propaganda ($M = 3.62$ on a 1–5 scale; 61.9% agreement). Confidence in identifying these phenomena was more mixed ($M = 2.95$; 38.1% agreement, 38.1% disagreement, 23.8% neutral).

When asked whether they would critically reflect on bias that supports their own worldview in the future, 61.9% responded *yes*, 28.6% *unsure*, and 9.5% *no*. As prior research repeatedly shows, self-assessment may diverge most from actual behavior in this area, people tend to recognize bias in others more readily than in themselves.

Participants who answered *yes* reported higher mean perceived learning ($M = 3.77$) and confidence ($M = 3.38$) compared to those who said *no* ($M = 3.50$ and $M = 2.50$, respectively). They also exhibited larger confidence gains ($\Delta M = +0.77$) and stronger reductions in perceived difficulty ($\Delta M = -1.46$), suggesting that reflective engagement was associated with greater perceived learning and ease of task performance.

Spearman correlations provided further insight into these relationships. Confidence improvement was strongly associated with perceived ease of the task ($\rho = -.711$; *confidence_delta* $\times$ *difficulty_delta*). Participants who reported higher confidence when supported also showed greater overall confidence gains ($\rho = .802$; *confidence_with* $\times$ *confidence_delta*). Learning and confidence were positively related, indicating that participants who felt they learned something new also tended to feel more confident afterward ($\rho = .550$; *learned_something_new* $\times$ *more_confident_after*).

In contrast, frequency of engaging with news content in one's daily life was negatively correlated with perceived learning ($\rho = -.564$; *learned_something_new* $\times$ *news_frequency_num*), suggesting that participants who frequently consumed news felt they learned less from the task. Finally, English proficiency showed moderate positive correlations with both baseline confidence ($\rho = .409$) and overall news consumption ($\rho = .509$).

5.2. Qualitative

In their feedback, participants repeatedly emphasized the difficulty of the task, particularly pointing to the large number of categories. This aligns with our intended task design, as we aimed for the task to feel challenging and impose a high cognitive workload.

This perception of difficulty is consistent with our survey results, where participants rated the task as relatively difficult on average.

Because of this perceived difficulty, some participants reported adopting a certain "sloppiness" in their approach, potentially not considering each category as thoroughly as would have been required for a perfect answer. Perfectionism did not seem realistically achievable under such demanding conditions. Participants also noted that this "sloppiness" increased with later sentences, suggesting that annotation quality may decline as the number of sentences grows. We anticipated this and therefore randomized sentence order and conditions to mitigate such effects.

Overall, these insights indicate that participants in a task of this kind may not sustain high levels

of effort, particularly if motivation is low. In our experiment, incentives were minimal: a small, unannounced raffle prize and, for some, interest in the topic itself. On crowdworking platforms such as Amazon Mechanical Turk, where participants are typically underpaid and compensated per sentence, incentives for sustained effort are even weaker, making the issue potentially more pronounced. We recommend hourly wages for production projects which is anyhow ethically merited.

6. Limitations

One might argue that by not using only native speakers of English, our results suffered from decreased quality. However, because each annotator completes both conditions, we argue that the setup is valid as potential effects would be visible in both conditions. Additionally, the aspect of cognitive load associated with annotation guidelines featuring more classes than one's short term memory may hold, is largely independent of one's native language.

7. Summary, Conclusion and Future Work

In this paper, we addressed the problem of annotating language resources for tasks with a large number of labels.

In order to improve annotation on these kinds of tasks, we developed *SALOMO*, which we evaluated in a controlled user study with 21 participants. The results show that LLM-assisted pre-selection substantially improved annotation efficiency and user experience. Participants completed items significantly faster (33.5 s on average), applied more and more diverse labels, and reported lower perceived difficulty and higher confidence in their annotations compared to the manual condition. Moreover, pre-selection guided annotators toward more stylistic and subjective bias categories, resulting in higher label consistency and stability ("stickiness"). This bias toward the pre-selection underscores the importance of a high-quality pre-selection process. Since the suggested labels strongly influence the final decision, they should be as accurate as possible from the outset.

Further feedback confirmed that, while the task remained cognitively demanding, participants appreciated the support and found it helpful for reflection on bias and media literacy. Overall, the findings indicate that LLM-assisted annotation can meaningfully reduce effort, increase label quality, and enhance user engagement in complex multi-label annotation scenarios.

At present, *SALOMO* is based on Web technology, but runs as a local application that generates annotated files. In future work, it could be integrated with a web server, so as to support collaborative work in a team and real-time reporting.

8. Ethics Statement

From an ethical point of view, one should distinguish between moral aspects involved in the annotation process and the moral implications of the result (in this case, the dataset and annotation tool).

As with any annotation project involving human subjects, they should be treated according to human experimentation best practices (ensuring informed consent/volunteer contribution, appropriate rewards and offering to debrief individuals after the experiment, if needed). We complied with these best practices.

Media bias and propaganda are geopolitically sensitive topics. As any dual good, a dataset for training the detection of media bias and propaganda might also be abused to serve as a tool to pass anti propaganda filters, or even to generate new instances of propaganda. We are not aware of a solution to this dilemma, but we hope that our work will serve the public good.

References

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024. Order effects in annotation tasks: Further evidence of annotation sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86, St Julians, Malta. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Max Bernhardt, Daniel C. Castro, Ryutaro Tanno, et al. 2022. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1):1161.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114.

Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S. Bernstein, Alex Berg, and Li Fei-Fei.

2014. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 3099–3102, New York, NY, USA. Association for Computing Machinery.

Bryan Guan, Tanya G. Roosta, Peyman Passban, and Mehdi Rezagholizadeh. 2025. The order effect: Investigating prompt sensitivity in closed-source llms. *arXiv preprint arXiv:2502.04134*. Accessed: 2025-09-10.

Austin Z. Henley and David Piorkowski. 2024. Supporting annotators with affordances for efficiently labeling conversational data. *arXiv preprint arXiv:2403.07762*.

HumanSignal. 2025. Label studio: Open source data labeling. Accessed: 2025-09-10.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176, St. Julians, Malta. Association for Computational Linguistics.

J. C. Klie, R. E. de Castilho, T. Neumann, S. G. Günter, and M. Gertz. 2018. The inception platform: Machine-assisted and knowledge-driven web annotation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1–10. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Thousand Oaks, CA.

Johann Laux, Fabian Stephany, and Alice Liefgreen. 2025. Better pay, clearer guidance: Investing in the working conditions of artificial intelligence data workers. *Big Data & Society*, 12(2):20539517251351320.

Jochen L. Leidner and Luca Jung. 2024. TAME II: A modern geographic text annotation tool. In *Proceedings of the 21st International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2024)*, volume 14673 of *Lecture Notes in Computer Science*, pages 95–104. Springer Nature.

Anna Sofia Lippolis, Mohammad Javad Saeedizade, Robin Keskisärkkä, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. 2025. Large language models assisting ontology evaluation. *arXiv preprint arXiv:2507.14552*.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. *arXiv preprint arXiv:2505.17510*.

Adrien Marchal, Merel C. J. Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3658–3671, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Winter Mason and Duncan J. Watts. 2010. Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, 11(2):100–108.

Tim Menzner and Jochen L. Leidner. 2026. The table of media bias elements: A sentence-level taxonomy of media bias types and propaganda techniques. *arXiv preprint arXiv:2601.05358*.

George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. In *Interspeech 2024*.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795, Vienna, Austria. Association for Computational Linguistics.

Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, Dakuo Wang, Lucy Yip, Liat Ein-Dor, Lena Dankin, Ilya Shnayderman, Ranit Aharonov, Yunyao Li, Naftali Liberman, Philip Levin Slesarev, Gwilym Newton, Shila Ofek-Koifman, Noam Slonim, and Yoav Katz. 2022. Label sleuth: From unlabeled text to a classifier in a few hours. Accessed: 2025-09-10.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, FL, USA. Association for Computational Linguistics.

Petter Törnberg. to appear. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, page 08944393241286471.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yifan Wang, David Stevens, Pranay Shah, Wenwen Jiang, Miao Liu, Xu Chen, Robert Kuo, Na Li, Boying Gong, Daniel Lee, Jiabo Hu, Ning Zhang, and Bob Kamma. 2024. Model-in-the-loop (milo): Accelerating multimodal ai data annotation with llms. *arXiv preprint arXiv:2409.10702*. Accessed: 2025-09-10.