

SAINE: Scientific Annotation and Inference Engine

Susie Xi Rao¹, Yilei Tu², Peter H. Egger³

¹ Chair of Applied Economics; Institute of Computing Platforms, ETH Zurich

² Chair of Applied Economics; Department of Computer Science, ETH Zurich

³ Chair of Applied Economics, ETH Zurich; CEPR; CESifo
{srao, yiletu, pegger}@ethz.ch

Abstract

We present SAINE, a Scientific Annotation and Inference ENgine based on a set of standard open-source software, such as Label Studio and MLflow. We show that our annotation engine supports the further development of our in-house hierarchical classification systems, in particular, of scientific documents. Based on our previous work on hierarchical discipline classifications, we demonstrate SAINE’s use for understanding the space for scholarly publications. The user study of our annotation results shows how the user input processed with SAINE can support a better understanding of the document classification process. SAINE can be particularly useful in fostering the transparency of meta-science research.

1 Introduction

A precise classification of publications across and within disciplines is key not only for a fast and comprehensive search to guide researchers to relevant material but also to identify the novelty of research, the standing and significance of scholars, and of the relative growth of fields of work (Bransford et al., 1999; Tsien, 2007; Kowsari et al., 2017; Frodeman et al., 2017; Semantic Scholar, 2022; Rao et al., 2023).

Machine learning develops into not only *a* but *the* customary approach to establish such a classification. Clearly, one would expect a search that is geared towards identifying a high-quality corpus of keywords to benefit crucially from supervision. Existing classifications of academic output, such as field of studies presented by Microsoft Academic Graphs (MAG) and OpenAlex (MAG successor), are based on a blend of (supervised) author-chosen and (unsupervised) machine-chosen keyword lists, where the composition of the blend is unknown to the researcher (Sinha et al., 2015; Shen et al., 2018; Wang et al., 2019a; Rao et al., 2023).

Prevailing systems, such as MAG, Dimensions, CrossRef, Web of Science (WoS), Scopus, or

Google Scholar, provide keywords for academic publications, which are lists based on abstracts in a discipline, field, and subfield, distilled from unsupervised machine learning, from word or phrase frequencies; or supervised learning, mostly from keyword self-reporting by authors; or semi-supervised learning, a mixture of the two, e.g., as done by MAG described in Sinha et al. (2015); Wang et al. (2019b, 2020). As reported in Rao et al. (2023), the quality of keywords in MAG hinders the further development of meta-science projects as their quality vary from discipline to discipline, and the granularity forbids us to have an even comparison of research output across disciplines, fields, and disciplines. Hence, the motivation of this work is two-fold. On one hand, it is of interest to the meta-science community to understand the generation of these keywords and how they are associated with the classification of disciplines, fields, and subfields. On the other hand, it is crucial to understand the outputs of the model in the classification systems (Rao et al., 2023) that we develop and continuously improve.

For designing an annotation and inference engine that helps establishing a classification system of scientific publications, one would target developing a tool with the following features: (1) a simple user interface with clear annotation instructions; (2) a reproducible pipeline across various disciplines; (3) good support for inference tailored to downstream tasks (e.g., model retraining) in meta-science studies. To the best of our knowledge, SAINE is the first scientific annotation engine specifically targeted towards understanding the model output in terms of label assignments and keyword representation. Our tool is one of its own kind in annotating scientific documents, which allows us to qualitatively understand an ML model in classifying scientific works.¹

¹We provide here a brief literature review for scientific annotations. Early work on understanding scientific documents

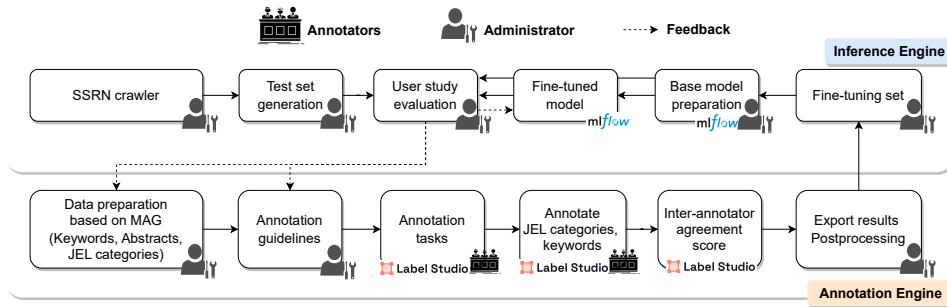


Figure 1: Saine Workflow and Pipeline.

Among the existing open-source annotation tools, Label Studio (Tkachenko et al., 2020-2022) suits those needs. Note that Gayoso-Cabada et al. (2019) have reviewed extensively the annotation tools that facilitate classification tasks. However, the reviewed tools are either not open-sourced or are domain-specific and, hence, do not share the aforementioned targeted features.

In this system demonstration, we utilize a set of standard open-source software, mainly Label Studio (Tkachenko et al., 2020-2022), MLflow,² FastAPI,³ and Pydantic⁴ to configure an annotation and inference engine for scientific publication annotations. In this demonstration, we illustrate the benefit of using supervised learning based on pre-established keyword lists and abstracts, and how annotators can help us better understand the importance of *supervised* learning in establishing a classification of academic publications.

This system is built on top of the hitherto largest-scale *multi-class* hierarchical classification study across all academic research disciplines in both *single-label* and *multi-label* settings (cf. Rao et al. (2023)). There, we have built a supervised hierarchical classification system that associates every

began with annotated segments of scientific articles for the purpose of information retrieval (Ibekwe-Sanjuan et al., 2011). Note that Gayoso-Cabada et al. (2019) have extensively reviewed the annotation tools that facilitate classification tasks. However, the reviewed tools are either not open-source or domain-specific, and hence do not share the aforementioned targeted features. Domain-specific efforts have been made to connect expert annotations and ontologies (Ciccarese et al., 2012). Ciccarese et al. (2011) discussed scientific document annotation as Resource Description Framework (RDF), which allows us to publish the annotation data as Linked Open Data, query it using languages such as SPARQL, and reason about it with SWRL. Small-scale studies (Zargayouna et al., 2016) have been conducted to annotate semantic roles in scientific articles. And knowledge graphs built on the topic of these annotations are studied (Luan et al., 2018; Zevio et al., 2020).

²<https://mlflow.org/>.

³<https://fastapi.tiangolo.com/>.

⁴<https://pydantic-docs.helpmanual.io/>.

publication with at least one and potentially several disciplines, fields, and subfields.

With the annotations above, we conduct a small user study with domain experts using our **annotation** engine. We then invoke our **inference** engine to fine-tune the base models in Rao et al. (2023). The comparison between the base and fine-tuned models shows that the proposed annotation and inference system is able to benefit the development of more accurate classifications.

To summarize, the paper presents a scientific annotation and inference engine called Saine, which is based on open-source software like Label Studio and MLflow. **The main contributions of the paper are:** (1) The demonstration of using Saine in understanding the space for scholarly publications, particularly in hierarchical discipline classifications. (2) The result of a user study, which shows that user input collected with the help of Saine can help better understand the classification process. (3) The ability of Saine to benefit the further development of a more accurate classification, demonstrated through the comparison between the base and fine-tuned models. (4) The potential of Saine to support downstream meta-science projects and foster greater transparency and understanding of scientific research.

The paper presents the benefits of supervised learning and the importance of a simple user interface with clear annotation instructions, reproducible pipelines, and good support for annotating and inferring scientific publications. The demonstration video can be accessed from <https://youtu.be/yTo0-G9YQK4>. A live demo website is available at <https://app.heartex.com/user/signup/?token=e2435a2f97449fa1> upon free registration. The project codebase is available at <https://gitlab.ethz.ch/raox/science-clf-/tree/labelstudio> and is merged with the codebase of Rao et al. (2023). We welcome collaboration

and feedback from the scientific community.⁵

In Figure 1 we illustrate the **workflow** in SAINE by assigning the roles of "Administrator", "Annotators", "Label Studio", and "MLflow" to each task in the pipeline. With Label Studio we assign the annotation tasks to annotators and curators; MLflow was used to keep track of pre-trained models reported in Rao et al. (2023), which makes it easy to load a certain model for inspection and inference by simply specifying the experiment ID tracked using MLflow. We then build the inference engine with FastAPI and Pydantic: The former can be used to deploy machine learning models as web API services; the latter can be used to define data models for input and output data, which helps prevent errors and improves the reliability of the API.

The sections are organized as follows. Section 2 introduces the functionality of Label Studio and its fit to our annotation needs, as well as our annotation guidelines for experts. Section 3 specifies the annotation design for the field of *Economics* and discusses the annotation results. Section 4 discusses the integration of annotation results into the pre-trained base models and the fine-tuned ones with MLflow. We devote Section 5 to discuss our preliminary experiments on improving annotation efficiency. We then conclude this system demonstration with a discussion of future work, system limitations, ethics, and broader impact statements.

2 Annotating Scientific Articles with Label Studio

We briefly introduce functionalities of Label Studio and illustrate why Label Studio is a suitable tool to annotate scientific articles. **The goals of annotations in the present context are three: (1) [To judge the appropriateness of an assigned category.]** Experts judge whether a scientific article has a correctly assigned category. If a category does not suit the abstract's content, the annotator will pick a new category from a pre-defined list. **(2) [To evaluate keywords.]** Experts evaluate keywords assigned by Microsoft Academic Graph (MAG) and mark missing keywords in the abstract. **(3) [To calculate inter-annotator agreement.]** The annotation engine should permit an efficient way of

calculating inter-annotator agreement (IAA) scores among annotators.

2.1 Label Studio

Label Studio is a powerful and versatile annotation tool that can handle various types of annotation tasks. Here are some of the features of Label Studio that make it suitable for the annotation tasks.

- **[Customizable interface.]** Label Studio allows us to design a customized annotation interface. One can create a pre-defined list of categories for the experts to choose from and provide them with clear instructions on how to evaluate the assigned category.
- **[Multiple annotation types.]** Label Studio supports various types of annotation, including text classification, entity recognition, and sequence labeling. Therefore, different types of annotation can be used to evaluate keywords, mark missing keywords, and judge predicted categories.
- **[Collaboration and version control.]** Label Studio enables multiple experts to work on the same project simultaneously, allowing efficient and collaborative annotation. It also includes a version control system that tracks changes to the annotations, facilitating easy comparison and IAA evaluation.
- **[Inter-annotator agreement (IAA).]** Label Studio has built-in tools to calculate IAA scores. These tools can help evaluate the consistency and reliability of chosen annotations.
- **[Integrations with machine-learning models.]** Label Studio also provides integration with various machine learning (ML) models. Although we do not use the integrated ML functions, Label Studio allows us to export the annotation results in JSON, with which we improve the classification models using the annotated data in the inference engine.

Overall, Label Studio offers a powerful and customizable annotation platform that can handle relevant annotation tasks, facilitate efficient collaboration among experts, and efficiently compute IAA.

The project manager uses an administrative panel (Figure 4 in Appendix B) to assign annotation tasks to each registered annotator and can monitor the annotation progress. The manager can also adjust the assigned annotations based on individual progress, as well as inspect tasks by annotation progress and IAA metrics.

⁵Note that the annotation engine is used for internal testing at the moment but will open for beta testing for all disciplines soon. We have over 100 users queuing for the beta stage. We are working on incorporating user profiles through self-declaration or automatic screening. The results of the beta stage will be shared with the community in due course.

2.2 Annotation Guidelines

When a publication is annotated, each annotator is provided with the abstract, the keywords offered by MAG, and the assigned category based on the keywords provided by MAG. The categories of a discipline classification (such as the Journal of Economic Literature, JEL, classification in economics) are assigned to MAG publications on the basis of the keywords. Therefore, MAG’s keywords help us identify potential misalignments and better understand the classifiers we built.

The annotation samples provided in the annotation engine are stratified sampled (ratio: 2e-5) across all classes of the training set introduced by Rao et al. (2023) for one discipline. Each annotator is required to judge whether a category is correctly assigned to an abstract. If not, the annotator is required to select the suitable one from a predefined list. The annotator is also required to evaluate MAG-generated keywords and make corrections (by removing unqualified keywords/markings suitable keywords from the abstract). Figure 2 shows two annotations of one publication. Label Studio makes it easy to navigate among the annotations generated by various annotators on an identical instance. Note that, as we discussed in Rao et al. (2023), our multi-class hierarchical classification system is modularized in both *single-label* and *multi-label* settings.⁶ The current annotation engine is equipped with both annotation functionalities. For the sake of system demonstration and user study in Section 3, we discuss the *single-label* setting. More details on the multi-label setting are provided in Appendix C.

3 Implementation: User Study in Economics

We now use *Economics* as a discipline to show how we utilize the annotation engine to collect expert annotations.

3.1 Annotation Design

We invited three economist experts from the Chair of Applied Economics at ETH Zurich to join the annotation project by accessing this link.⁷ Annotation guidelines are given at our project repository.⁸

⁶We offer an brief overview of our hierarchical classification system in Appendix A.

⁷<https://app.heartex.com/user/signup/?token=edd1c85b0a9c9cd2>.

⁸[https://gitlab.ethz.ch/raox/science-clf/-/blob/labelstudio/labelstudio/annotation_](https://gitlab.ethz.ch/raox/science-clf/-/blob/labelstudio/labelstudio/annotation_guideline.md)

Of the three experts, one has annotated all provided instances (Annotator 1), one has annotated 10% of the instances (Annotator 2), and one has annotated a subset of instances with an *ex ante* denomination in Urban and Spatial Economics only (Annotator 3). Each annotator received a user panel like Figure 5 in Appendix B.

3.2 Annotation Results in Label Studio

Altogether, 788 instances of abstracts and keywords from MAG had to be annotated for a *single-label* classification. In Economics, a standardized field and subfield system with keywords exists, and it is called the Journal of Economic Literature (JEL) classification system.⁹ This system is known to all academic economists and serves as a guiding principle to associate an article or a topic with a specific subfield in *Economics*. The subfields in the JEL categories are associated with keywords.

We report the annotation time and IAA scores across all tasks that are automatically calculated by Label Studio.¹⁰ The final task agreement score is calculated by averaging all IAA scores for each annotation pair. Table 1 illustrates the IAA scores amongst three experts. Annotators 1, 2, and 3 have annotated 788, 181, and 99 instances, respectively. The number of overlapping tasks between annotator pairs is 52 (Annotators 2 and 3), 99 (Annotators 1 and 3), and 181 (Annotators 1 and 2), respectively. The median annotation time of Annotators 1-3 per instance was 17.7s, 29.8s, and 40.9s, respectively. The annotators were entitled to disapprove of the assigned category based on MAG upon suggesting an alternative category. Marking and filling in missing keywords is time consuming, and reading the MAG-generated keywords can help, to some extent, the annotation speed. However, all annotators reported that MAG-provided keywords could be a source of error for incorrectly assigned categories. As discussed among the annotators after they underwent the annotations separately, the category they found the best was for *Mathematical & Quantitative Methods*, and it was worst for *Macroeconomics and Public Economics*.¹¹

[guideline.md](https://gitlab.ethz.ch/raox/science-clf/-/blob/labelstudio/labelstudio/annotation_guideline.md).

⁹<https://www.aeaweb.org/econlit/jelCodes.php?view=jel>.

¹⁰See the official documentation under this link (<https://labelstud.io/guide/stats.html>) for the IAA calculation steps. Using the “Basic matching function” gives us the current results in Table 1 while using the “exact matching” gives us even better results.

¹¹We are aware that we need to assign budget to incentive potential expert annotators. In collaboration with Open

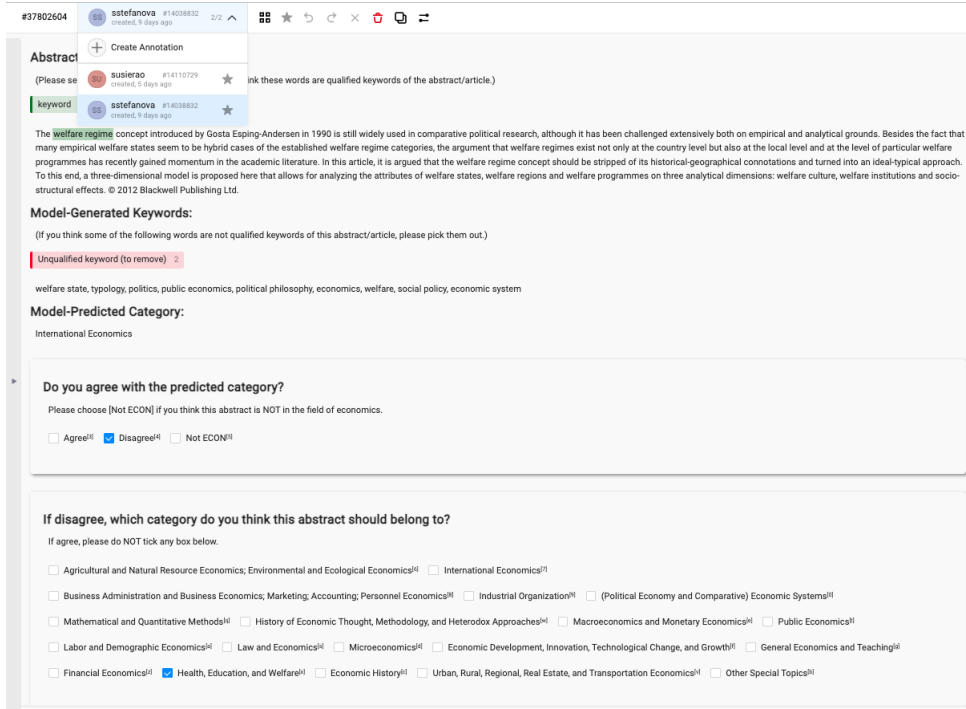


Figure 2: Publication Annotations by Multiple Annotators.

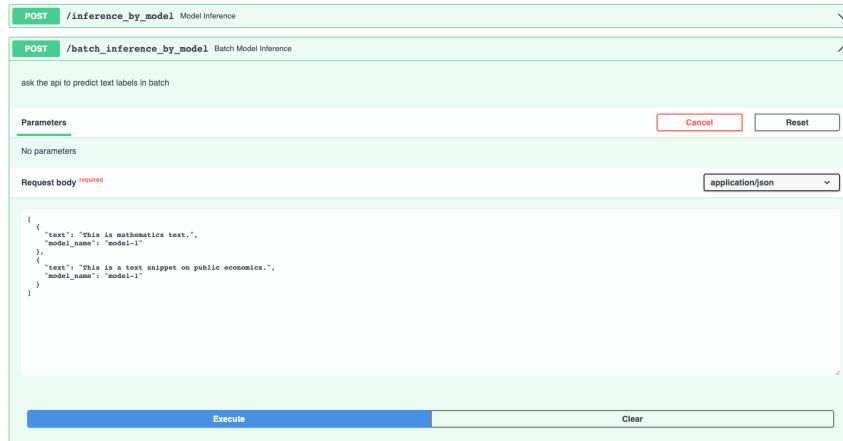


Figure 3: Inference Engine with *MLflow* Integration. API: Batch Inference by Model.

| | Annotator 1 | Annotator 2 | Annotator 3 |
|-------------|-------------|-------------|-------------|
| Annotator 1 | | 55% | 55% |
| Annotator 2 | 55% | | 27% |
| Annotator 3 | 55% | 27% | |

Table 1: Annotator Agreement Matrix Among Three Expert Annotators.

4 Inference Engine: Incorporating Annotation Results into the Existing Classification Pipeline

We illustrate the pipeline using the discipline of *Economics* as discussed in Section 3.

Research Data Call (<https://ethrat.ch/en/eth-domain/open-research-data/>) and ETH Library, we are exploring possible ways to remunerate high-quality annotations.

4.1 Post-processing of Annotation Results

We downloaded the annotation results in JSON¹² of all experts and post-processed them following the protocols below, before feeding them into the pre-trained base models of various neural networks as discussed in Rao et al. (2023). In total, we obtained 1,068 partly overlapping annotations (incl. "Skip", "(Dis)agree", keywords, added categories). The basic statistics on the number of instances of "Agree", "Disagree" and "Not ECON" are 498, 297, and 268,

¹²The annotation results are accessible under https://gitlab.ethz.ch/raox/science-clf/-/blob/labelstudio/labelstudio/labeled_data/labelstudio_econ.json.

respectively.

The post-processing procedure is structured as follows. (1) We removed abstracts that were inadequately classified as belonging in *Economics* from the sample (206 of 788 instances). Additionally, we deleted 5 instances due to bad annotations. For example, no one labeled this sample ("Skip"), or an annotator chose "Disagree" but did not choose a new category. (2) For each remaining instance, we counted the percentages of "Agree" and "Disagree" verdicts relative to the label generated on the basis of MAG keywords. If strictly more experts agreed than disagreed with MAG, the original label was preserved (for 351 of the 577 valid instances). Otherwise, we took the label suggested by the majority of annotating experts (for 226 of the 577 valid instances). (3) In the case of ties, we randomly picked a label from the suggested annotations (for 22 of the 226 category-renewed instances). Following this protocol, we obtained 561 instances with expert-curated labels to fine-tune the base models.

4.2 Fine-tuning Pre-trained Base Models

We used the 561 labels generated by the experts as a fine-tuning set on the base models reported in Rao et al. (2023) on the discipline of *Economics* (model-1). We compared the inference performances of the base model (**Model** in Table 2) with those of the fine-tuned model (**Model_FT** in Table 2) on various neural network architectures, Deep Neural Network (DNN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Transformers. To benchmark the differences in performances between **Model** and **Model_FT**, we created a small test set from the Social Science Research Network (SSRN),¹³ which is a website that provides a platform for researchers to share and distribute their research papers and other scholarly work in the social sciences and other related fields. We decided to use the *Economics* SSRN publications because they come with human-curated JEL categories, keywords, and abstracts.

Concretely, we built an SSRN crawler¹⁴ to download the publication space in *Economics* publications in SSRN, where all contained research articles

¹³The test set is accessible under <https://www.ssrn.com/index.cfm/en/>.

¹⁴Our crawler is accessible under https://gitlab.ethz.ch/raox/science-clf/-/tree/labelstudio/labelstudio/ssrn_crawler and the publications on economics on SSRN can be retrieved under <https://papers.ssrn.com/sol3/displayjel.cfm>.

| Model (1) | Model_FT (2) | $\Delta(\text{Model_FT}-\text{Model})$ (3) | Model = Model_FT (4) | Total (5) | Model Type (6) |
|--------------|-----------------|--|-------------------------|--------------|-------------------|
| 54 | 58 | 4 | 108 | | CNN |
| 59 | 69 | 10 | 148 | 190 | RNN |
| 39 | 39 | 0 | 190 | | DNN |
| 31 | 33 | 2 | 37 | | Transformer |

Table 2: Results of the User Study. *FT*: Fine-tuned.

in *Economics* are multi-category-indexed. This means, each publication there is indexed by at least one JEL code and it allows multiple JEL codes per publication. We could easily validate with our multi-label engine in principal, but we focus on *single-label* classifications for this user study.

To create this test set, we randomly sampled 10 instances from each of 19 JEL field classes, which resulted in a sample of 190 test instances. In the implementation of hierarchical classifications reported in Rao et al. (2023), we have used MLflow to track and manage ML experiments, with which we have saved all pre-trained base models. Now, based on them, we could seamlessly integrate model fine-tuning and inference with various models. The inference engine API has been implemented using FastAPI with help from Pydantic.¹⁵ We illustrate the batch inference API in Figure 3, with which users can feed the test set into various models (base or fine-tuned) and obtain predictions. In Appendix D we provide more details about the inference engine.

4.3 Benefits of Expert Annotations

We present the results of user studies in Table 2. Specifically, we inspect two types of statistics, the correct predictions of the base and fine-tuned models in Columns (1)-(2), and the identical predictions of the base and fine-tuned models in Column (4). Since each publication is multi-JEL-category-indexed, we count the prediction as "correct" if the indices include the predicted category. Column (1) is the base model trained with the model type specified in Column (6). Column (2) presents the results of the fine-tuned (supervised) model. Column (4) shows that out of a total of 190 test instances, identical predictions were generated by the base and fine-tuned models. We see that fine-tuning with user-generated results has brought benefits to all

¹⁵FastAPI (<https://fastapi.tiangolo.com/>) and Pydantic (<https://pydantic-docs.helpmanual.io/>) are two Python libraries commonly used in machine learning and web development to build robust and efficient APIs. Combining them allows us to create APIs with strong data validation, asynchronous support for handling multiple requests, and automatic documentation generation. This combination simplifies the deployment of machine learning models as web services while maintaining code quality and reliability.

models except DNN because DNN predicts for all test examples only one class (the dominating one). RNN is the best performer when considering the benefits resulting from expert supervision, because the Δ in correct predictions has increased the most according to Column (3). Interestingly, fine-tuning a pre-trained Transformer model may not always result in a significant improvement in performance, as we see from a comparison with other base models. However, the current fine-tuning set is too small to draw firm conclusions in this regard.

5 Discussions: Improving Annotation Efficiency

We share preliminary results of improving annotation efficiency based on the annotators’ feedback.

5.1 Similarity between Articles and Scholars

First, we assume that we have access to the annotators’ Google Scholar profiles and their self-reported fields of expertise.¹⁶ Then we assume that these two sets of labels do not vary largely, which gives us a high confidence score about one annotator. Finally, we match the best-suited scholars to the articles to annotate by extracting keywords from the top-cited articles of these scholars and scoring them on the cosine similarity with the article keywords. The results are promising and can reduce work overhead for the scholars while improving the quality of the annotations by assigning best suited scholars for the process. More details on the implementation are given in Appendix E.

5.2 LLM as Annotators

Do we still need the supervised learning models now that we have the mighty LLM models? How well do LLMs perform these annotation tasks? In light of our commitment to total project transparency, we have opted to utilize the Vicuna 7B and 13B models (Chiang et al., 2023), both of which are publicly available for non-commercial use and are fine-tuned based on LLaMA (Touvron et al., 2023), explicitly tailored for QA tasks. For their predictions on single-label and multi-label classifications, see Tables 3 and 4. Overall, we observe that even one of the best LLMs performs poorly in the *single-label* setting and has potential to be

¹⁶Each annotator will be asked to fill in a self-declaration form (https://gitlab.ethz.ch/raox/science-clf/-/blob/labelstudio/labelstudio/Scholar%20Similarity/Researcher_Details.pdf) before we assign the annotator’s tasks.

| Vicuna | Agree | Disagree | NOT ECON | Blank | Total |
|--------|-------|----------|----------|-------|-------|
| 7B | 740 | 28 | 16 | 4 | 788 |
| 13B | 122 | 648 | 7 | 11 | 788 |

Table 3: Results of *Single-Label* Prediction by Vicuna. Those Vicuna outputs that do not contain the keywords "Agree", "Disagree" or "NOT ECON" are labeled as "blank".

| Vicuna | ECON (Y/N) | Cat1 (A/D/B) | Cat2 (A/D/B) | Cat3 (A/D/B) |
|--------|------------|-----------------|----------------|--------------|
| 7B | 417 / 297 | 157 / 146 / 114 | 122 / 199 / 96 | 5 / 18 / 10 |
| 13B | 614 / 100 | 158 / 445 / 11 | 121 / 490 / 3 | 3 / 39 / 0 |

Table 4: Results of *Multi-Label* Prediction by Vicuna. Y/N - "Yes"/"No", A/D/B - "Agree"/"Disagree"/"Blank". "Cat1,2,3" are three predicted labels by our classification system. If it predicts that an abstract does not belong to ECON, then we will no longer ask whether or not it agrees with the our model-predicted categories. Responses that do not contain the specified keywords are considered "blank". The dataset contains a total of 42 entries with non-empty Cat3. In this subset, Vicuna-13B uniformly classifies all entries to be in the ECON domain. Conversely, Vicuna-7B predicts 33 of these entries within the ECON domain.

used as keyword extractor and annotate *multi-label* classifications. The details of the experimental protocols are in Appendix F.

6 Conclusions and Future Work

In this system demonstration, we utilize a set of standard open-source software, mainly Label Studio (Tkachenko et al., 2020-2022), MLflow, FastAPI and Pydantic, to configure an annotation and inference engine for scientific publications (SAINE). This system is built on top of hitherto largest multi-class hierarchical classification study across all disciplines in both single-label and multi-label settings (cf. Rao et al. (2023)). We illustrate the functionality of the system with a user study in *Economics* and show that expert inputs to our system can help better understand the classification process, which benefits the development of a stronger model in the next iteration. We plan to open-source the data and codebase and invite collaborative work in the direction of meta-science. SAINE can bring about societal and scholarly impact in enabling more transparent and rigorous scientific research understanding by promoting ethical AI usage and responsible research practices.

In future work, we plan to expand SAINE to include different scientific disciplines, which allows us to conduct larger-scale user studies; we also plan to explore *multi-label* annotation tasks in all levels including disciplines, fields and subfields. As already discussed in Section 5, we plan to leverage LLMs to assist in annotations, as well as to automatically select the suitable annotators based on their profiles and publications to annotate.

Limitations

Label Studio has some limitations in incorporating existing ML pipelines into the annotation engine, especially, when using customary code. It is not clear from the documentation (<https://labelstud.io/guide/ml.html>) how to deploy the ML backend with pre-trained models. We will discuss this with the developers at Label Studio and see how we can bring the annotation engine and the ML pipeline closer to each other.

In terms of annotator selection, at the moment we have to select the experts for each discipline. However, we have performed experiments to rank annotators according to their field expertise and find the best annotation tasks based on the similarity between the space for academic publications and the space for articles (Appendix E). One future idea is to automatically compute an associative score between a third-party academic product such as Google Scholar and the publication space. For instance, the project PeopleMap¹⁷ provides interesting techniques to generate researcher profiles based on their research interests and publications taking as input the URLs of the researchers' Google Scholar profile. At this stage, Label Studio developers suggest that we add a self-declarative questionnaire to each annotator, which can be used as meta-data on annotators when quantifying the annotation confidence score. Due to time constraints, we have not yet added this questionnaire in the user study reported in the current manuscript, as the experts in the current user study are selected by our project PI and have strong expertise in *Economics*. However, we will incorporate this self-declaration questionnaire in the beta testing stage.

In terms of annotation efforts, we have benchmarked annotation quality using LLMs, which shows that human annotators are needed to control the quality. Considering our annotators' feedback that it is time-consuming to extract keywords for humans, it makes sense to use LLMs as an annotation-assisting engine for keyword extraction at this stage. We have evaluated the LLM-generated keywords: some are quite generic given the context, and others are good fits. We plan to do a systematic evaluation of LLM-generated keywords using the WOS-46985 benchmark dataset.¹⁸ In terms of label prediction, we see in Tables 3 and

4 that Vicuna performs poorly on the *single-label* task and we will need to finish a larger sample of *multi-label* task to gauge its values, despite its superior performances in 100 tasks we evaluate against human performance.

Ethics Statement

We acknowledge that our system may involve processing potentially sensitive data (such as annotator profile), and we take data privacy and ethical considerations very seriously. In accordance with ethical guidelines of "ACM Code of Ethics", we will take steps to protect the privacy of annotators once the annotation engine is in the beta stage. Only annotation administrators have access to the results of researcher questionnaires. On the Label Studio platform (see <https://labelstud.io/blog/human-feedback-in-ai/>), the annotators' tasks and profiles are unknown to their peers. We also enable a comparison between subjective evaluation (by self-declaration) and objective evaluation (by past publication records of the annotators) to mitigate biases. We have also made efforts to ensure that our system and its annotations are unbiased and fair. We believe that our work will help foster greater transparency and understanding in scientific research, and we welcome collaboration and feedback from the scientific community to further advance ethical and responsible use of AI in research.

Broader Impact Statement

Our annotation engine and inference engine can further support downstream meta-science projects. We list a few interesting questions we can answer using our pipeline (Rao et al. (2023) and the annotation and inference engine).

1. **[For students.]** Which fields of research are more impactful/growing?
2. **[For policy makers.]** How to design education for cross-/inter-/pluridisciplinary studies?
3. **[For department and tenure committees.]** How to benchmark output and impact levels of an untenured scholar?
4. **[For funding institutions.]** How to measure/quantify inter-/pluri-disciplinary standards for institutions such as SNIS and SNSF that emphasize the interdisciplinarity of research?¹⁹

¹⁷<https://github.com/poloclub/people-map>.

¹⁸<https://paperswithcode.com/sota/document-classification-on-wos-46985>.

¹⁹SNIS: <https://snis.ch/>, SNSF: <https://www.snf.ch/en>.

5. **[For librarians.]** How can one effectively organize bibliographical resources across disciplines and departments in one university?²⁰

We plan to add other disciplines covered by Rao et al. (2023) to our annotation engine. We would also like to incorporate subjective (self-declaration) and objective measurements (e.g., Google Scholar profile integration) into the annotation pipeline. This may help develop confidence scores of one annotation/annotator.

Acknowledgements

We thank the colleagues at DS3Lab (<https://ds3lab.inf.ethz.ch/>) for providing valuable feedback when prototyping the system design. Without the strong support of our expert group at the Chair of Applied Economics (<https://cae.ethz.ch/>), the user study would not have been possible. We thank Mr. Prakhar Bhandari and Ms. Piriya Korn Piriya Tamwong for their technical support to our project. We appreciate that Label Studio has offered us an academic license for the project, which allows us to invite more experts to contribute in the long run. The user agreement and terms of an academic license are listed <https://labelstud.io/academic/>.

References

- John D Bransford, Ann L Brown, and Rodney R Cocking. 1999. How people learn: Brain, mind, experience, and school. *Washington, DC: National Academy Press. ED*, 436:276.
- Marco Campos. 2020. Yake: Yet another keyword extractor. <https://github.com/LIAAD/yake>.
- Sahil Chaddha. 2020. Python implementation of rapid automatic keyword extraction using nltk. https://csurfer.github.io/rake-nltk/_build/html/index.html.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Paolo Ciccarese, Marco Ocana, and Tim Clark. 2012. Open semantic annotation of scientific publications using DOME0. In *Journal of Biomedical Semantics*, volume 3, pages 1–14. BioMed Central.
- Paolo Ciccarese, Marco Ocana, Leyla Jael Garcia Castro, Sudeshna Das, and Tim Clark. 2011. An open annotation ontology for science on Web 3.0. In *Journal of Biomedical Semantics*, volume 2, pages 1–24. BioMed Central.
- Robert Frodeman, Julie Thompson Klein, and Roberto Carlos Dos Santos Pacheco. 2017. *The Oxford Handbook of Interdisciplinarity*. Oxford University Press.
- Joaquín Gayoso-Cabada, Antonio Sarasa-Cabezuelo, and José-Luis Sierra-Rodríguez. 2019. A review of annotation classification tools in the educational domain. *Open Computer Science*, 9(1):299–307.
- Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.
- Fidelia Ibekwe-Sanjuan, Fernandez Silvia, Sanjuan Eric, and Charton Eric. 2011. Annotation of scientific summaries for information retrieval. *arXiv preprint arXiv:1110.5722*.
- Kamran Kowsari, Donald E Brown, Mojtaba Heidarzafar, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. HDLTex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371. IEEE.
- Yi Luan, Luheng He, Mari Ostendorf, and Hananeh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.
- OpenAI. 2023. Gpt-4 technical report.
- Susie Xi Rao, Peter H Egger, and Ce Zhang. 2023. Hierarchical classification of research fields in the "web of science" using deep learning. *arXiv preprint arXiv:2302.00390*.
- Susie Xi Rao, Piriya Korn Piriya Tamwong, Parijat Ghoshal, Sara Nasirian, Emmanuel de Salis, Sandra Mitrović, Michael Wechner, Vanya Brucker, Peter Egger, and Ce Zhang. 2022. Keyword extraction in scientific documents. *arXiv preprint arXiv:2207.01888*.
- Semantic Scholar. 2022. Announcing s2fos, an open source academic field of study classifier. *AI2 Blog*. [Online; accessed: 2022-11-22].
- Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. *arXiv preprint arXiv:1805.12216*.
- Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darin Eide, Bo-june Paul Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM.

²⁰Our team is invited to contribute to a blog post of Swiss Year of Scientometrics (<https://yearofscientometrics.ethz.ch/>), a project of ETH Library, funded by swissuniversities (<https://www.swissuniversities.ch/en/>).

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. [Label Studio: Data labeling software](https://github.com/heartexlabs/label-studio). Open source software available from <https://github.com/heartexlabs/label-studio>.

TogetherComputer. 2023. [OpenChatKit: An Open Toolkit and Base Model for Dialogue-style Applications](#).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).

Joe Z Tsien. 2007. The memory code. *Scientific American*, 297(1):52–59.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019a. [A review of Microsoft Academic Services for science of science studies](#). *Frontiers in Big Data*, 2:45.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019b. [A review of microsoft academic services for science of science studies](#). *Frontiers in Big Data*, 2.

Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *LREC 2016*.

Stella Zevio, Guillaume Santini, Henry Soldano, Haïfa Zargayouna, and Thierry Charnois. 2020. A combination of semantic annotation and graph mining for expert finding in scholarly data. In *Proceedings of the Graph Embedding and Mining (GEM) Workshop at ECML PKDD*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).

A Our Hierarchical Classification System

We provide an overview of [Rao et al. \(2023\)](#). The presented paper introduces a modularized three-level hierarchical classification system designed to automatically categorize scholarly publications based on their abstracts. The system operates

within a hierarchical label set consisting of disciplines, fields, and subfields, enabling multi-class classification. This approach facilitates a systematic categorization of research activities, considering both knowledge production and impact through citations. The system distinguishes 44 disciplines, 718 fields, and 1,485 subfields, leveraging a vast collection of abstract snippets from the Microsoft Academic Graph. By utilizing various neural network models, such as DNNs, RNNs (using GRU), CNNs, and Transformers, through batch training, the system achieves high classification accuracy rates exceeding 90% in both single-label and multi-label settings.

The modular design of the system allows for flexibility and easy integration of new models, with CNNs identified as the most efficient performer across the models. The system consists of three components: the first component (L1) handles discipline classification, the second component (L2) focuses on field classification, and the third component (L3) specializes in subfield classification. Each component operates based on the output of the previous level, enabling a granular categorization of research activities and capturing the interdisciplinary nature of certain topics.

In the classification process, the system assigns publications to disciplines, fields, and subfields based on their abstracts. It computes conditional probabilities to determine the relevance of each label given the previous level labels. This hierarchical approach improves the alignment of research texts with disciplines, enables automated classification, and captures interdisciplinarity.

The system incorporates both single-label and multi-label settings. In the single-label setting, each publication is assigned to a single category, while in the multi-label setting, publications can be assigned to multiple categories simultaneously. The multi-label classification assumes label independence and employs binary cross-entropy loss for training. To ensure a balanced distribution of relevant and irrelevant samples, stratified sampling is maintained for label sets.

Performance evaluation of the classification system includes metrics such as categorical accuracy, precision, and recall. The system’s ability to accurately classify research texts, align them with relevant disciplines, and capture interdisciplinarity contributes to its value in indexing and analyzing scientific publications.

Overall, the proposed system, with its modular design and pretrained models, serves as a solid foundation for future applications in scientific publication indexing and analysis.

B Label Studio Functionalities

In Figure 4 we demonstrate the administrative panel of the project manager. The "Filters" and "Order (Annotation results)" tabs make it easy to inspect tasks by annotation progress (e.g., "Annotators", "Agreement", "Completed", "Total annotations per task"). In Figure 5, we demonstrate the user panel that we show for each expert annotator. Each annotator has no access to additional information about the annotations made by the other annotators. As an expert, one can only see how many annotations have been gathered per instance among the experts together.

C Multi-label Annotation Engine

The setup of multi-label annotation is similar to the single-label setting. In the multi-label settings, annotators are required to mark "(Dis)agree" for each suggested JEL category (we provide three categories at most) and then select additional JEL categories, where multiple choices are allowed.

D Inference Engine

We provide two types of API calls for inference, (1) `inference_by_model` and (2) `batch_inference_by_model`. The only difference between these two API calls is that API (2) allows text-label predictions in batches, which requires a JSON sequence as input. Figure 7 demonstrates the user interface.

E Similarity between Authors and Articles

E.1 Keyword Extraction

The process of keyword extraction from the top 25 cited articles of each author and individual publication articles involves the following steps and methods. The scholars we picked are five renowned economists and use Google Scholar to download their profiles and publications. We start with the abstracts from these articles as the primary source of information.

Firstly, we perform an initial cleaning and pre-processing on these abstracts. This cleaning involves the removal of non-alphanumeric characters,

conversion of text to lowercase, and tokenization of the text into individual words. We also remove common words, known as stop words, which do not contribute much to the overall meaning of the text. Finally, we discard words that are less than three characters long as these are typically not meaningful. This cleaning process results in a simplified and standardized version of the original text which is more suitable for further analysis.

Secondly, we ensure that all our data is in English to maintain consistency. For this, we employ a language detection function. If a text is not in English, we translate it using a translation pipeline, which is a model capable of accurately translating text from various languages to English. To handle potential memory issues with larger texts, we split the text into smaller chunks, translate each chunk separately, and then concatenate them back together.

The cleaned and translated text is then passed through KeyBERT (Grootendorst, 2020), a minimalistic transformer-based keyphrase extraction technique, which is used to extract keywords from the text. Apart from KeyBERT we also tried other keyword extraction techniques like YAKE (Campos, 2020) and RAKE (Chaddha, 2020). We then compared which of these techniques extracted the best keywords by generating scores for the tasks on the benchmark data set "WOS-46985" discussed in (Rao et al., 2022). We found that KeyBERT got the best scores for the extracted keywords when compared to the reference data. KeyBERT uses BERT, a state-of-the-art transformer model for natural language processing, to convert words into high-dimensional vectors or embeddings. These embeddings capture the semantic meaning of the words and their context. KeyBERT then identifies clusters in these embeddings to find the most representative or "key" phrases.

We extract 250 keywords for each author by combining and analyzing the abstracts of their top 25 cited articles, which gives us a broad representation of their research interests. For individual publication articles, we extract 15 keywords to capture the essence of each specific article. We have tried 5, 10, 15 keywords for each article, and 15 gives the best results.

By following this methodology, we ensure the extraction of the most relevant and informative keywords for each author and individual article, providing us with a valuable understanding of the research

| ID | Completed | Date | Annotators | Agreement | Reviewers | cat1 | Keywords | abstract | PaperID |
|----------|--------------------------|-----------------------|------------|-----------|-----------|---|---|---|------------|
| 37802194 | <input type="checkbox"/> | | 1 | 100% | | Agricultural and Natural Resource Economics & Environmental and | casting, water cooling, materials science, porosity, composite material, | Aimed at the problem that water cooling channel and core in heavy sectioned | 2353936017 |
| 37802788 | <input type="checkbox"/> | Feb 13 2023, 15:11:21 | 2 | 100% | | Public Economics | electrical network, recursion, real time computing, data validation, | Abstract Based on the PROLDG programming language, the paper | 2014825585 |
| 37802705 | <input type="checkbox"/> | | 2 | 0% | | Macroeconomics and Monetary Economics | fem, grassland, shrubland, shrub, primary production, m 2, vegetation, biomass, | Carbon dynamics of the grass, fern and shrub communities in Hong Kong | 2372699118 |
| 37802278 | <input type="checkbox"/> | | 1 | 100% | | Business Administration and Business Economics & Marketing & | engineering ethics, subtitle, automotive engineering, advertising, engineering | SUBTITLE: NEGATIVE FACTORS SEEM TO BE GAINING THE UPPER | 620453483 |
| 37802011 | <input type="checkbox"/> | Feb 09 2023, 13:41:59 | 2 | 0% | | Agricultural and Natural Resource Economics & Environmental and | feature vector, spectroradiometer, land cover, geology, decision | Abstract Several investigations indicate that the Bidirectional | 2032180438 |
| 37802706 | <input type="checkbox"/> | Feb 13 2023, 15:05:47 | 2 | 100% | | Macroeconomics and Monetary Economics | formic acid, carbon, bimetal, catalysis, nanoparticle, metal organic | Abstract An efficient synergistic catalyst, AgPd bimetal supported on | 2257321918 |
| 37802042 | <input type="checkbox"/> | Feb 13 2023, 11:15:24 | 2 | 0% | | Agricultural and Natural Resource Economics & Environmental and | agriculture, moisture, ulitsoi, insectioi, andioi, irrigation, environmental | Abstract The district of Central Aceh is one of district located in the | 2593276985 |
| 37802279 | <input type="checkbox"/> | Feb 14 2023, 13:58:10 | 2 | 100% | | Business Administration and Business Economics & Marketing & | citrate synthase, protease, proteases, endurance training, skeletal muscle, | Alkaline and myofibrillar protease activities of rectus femoris, soleus, and tibialis | 2070676317 |
| 37802516 | <input type="checkbox"/> | | 1 | 100% | | Industrial Organization | telecommunications link, electronic engineering, throughput, diversity gain, | We introduce a distributed opportunistic scheduling (DOS) strategy based on | 1967382014 |
| 37802184 | <input type="checkbox"/> | | 1 | 100% | | Agricultural and Natural Resource Economics & Environmental and | troposphere, climatology, anticyclone, geology, meteorology, precipitation, | The relationship between 10–30 d low-frequency oscillation and the rainfall | 2360107436 |
| 37802150 | <input type="checkbox"/> | Feb 14 2023, 13:30:56 | 2 | 0% | | Agricultural and Natural Resource Economics & Environmental and | precautionary principle, indexation, hazard analysis, water resource | Abstract Wastewater management receives a great deal of attention with | 1981386852 |

Figure 4: Administrative Panel of Annotation Tasks in Label Studio.

| ID | cat1 | Keywords | abstract | PaperID |
|----------|--|--|--|------------|
| 37802242 | Business Administration and Business Economics & Marketing & | advertising, sociology, performance art | La mayoría de las investigaciones sobre la publicidad de la televisión | 2148438171 |
| 37802293 | Business Administration and Business Economics & Marketing & | content analysis, cultural center, advertising, exhibition, architecture, | First of all demanded changes in cultural politics are connected with the | 335718745 |
| 37802295 | Business Administration and Business Economics & Marketing & | chauvinism, public relations, advertising campaign, dozen, german, girl, desk, | r~ w will require egional, communication strategies as consUlmer preferences | 1964706606 |
| 37802302 | Business Administration and Business Economics & Marketing & | trip generation, precinct, pedestrian, advertising, town centre, city centre, | In autumn 1981 the community of Amersfoort (about 90,000 inh.) initiated | 618293861 |
| 37802303 | Business Administration and Business Economics & Marketing & | cycling, endurance training, overweight, physical therapy, vo2 max, | The effect of different doses of endurance training on the capacity to oxidize | 1531576420 |
| 37802304 | Business Administration and Business Economics & Marketing & | supply and demand, marketing, business, operations management | The evolutionary development of a global logistics strategy in ICL is | 1989919588 |
| 37802305 | Business Administration and Business Economics & Marketing & | product innovation, marketing management, digital marketing, marketing | A considerable number of studies have been assembled over the last | 2013203950 |

Figure 5: Annotator Panel of the Assigned Annotation Tasks in Label Studio.

landscape and the interests of the authors.

E.2 Similarity Scores

In our work, we present an innovative method that allows for a comprehensive understanding of the relationship between authors, publications, and research categories. This approach uses a function which not only identifies the top authors relevant to a particular article but also uncovers the top fields of research or "categories" connected to the article and hence its best annotator(s).

The function employs the TF-IDF (Term Frequency-Inverse Document Frequency) vector-

ization technique to transform text data into a numerical representation that can be processed by machine learning algorithms. For a given publication, it uses this technique to compare the article's abstract to those of top-cited authors, generating a list of the most similar authors.

Subsequently, the function identifies the top research categories linked to the publication by analyzing the keywords in its abstract. It applies the same process to the top-ranked author's 25 most cited articles. The result is a set of top categories that best align with the publication and the most relevant author, providing a deeper understanding

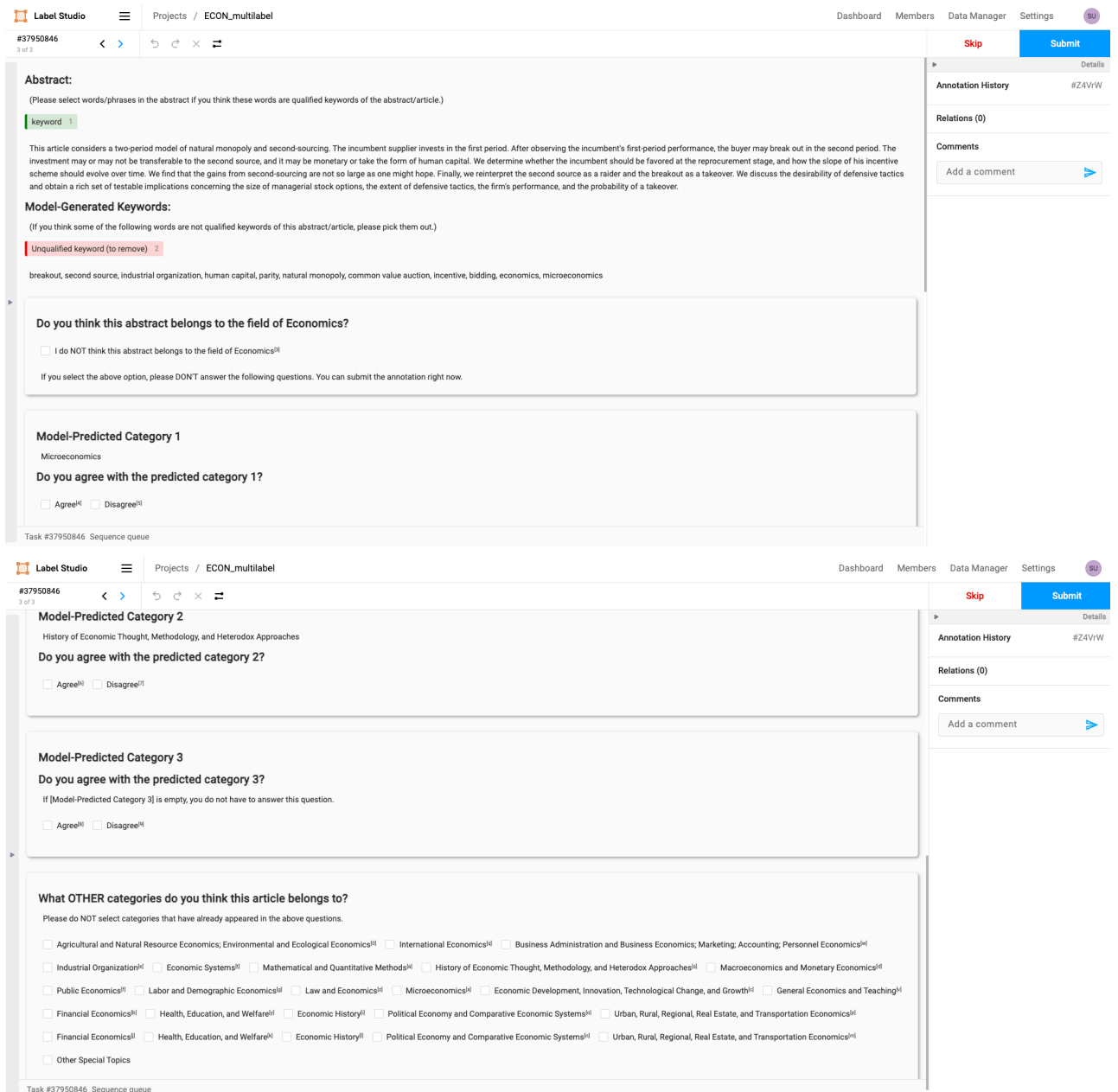


Figure 6: Publication Annotation Engine in a Multi-label Setting.

of their research focus. This novel approach offers a multidimensional view of the research landscape, establishing clear links between authors, their publications, and research fields.

E.3 Plotting the Author-Article Similarities

In our research, we have developed a method for visualizing the semantic proximity between a specific publication and the top 25 cited author publications across all authors. This is accomplished through a function that maps the abstracts of the documents into a two-dimensional space using Word2Vec for word embeddings and PCA for dimensionality re-

duction. The resulting plot provides a graphical representation of how closely related the content of a given publication is to the influential works of various authors. In Figure 8 we show an example plot produced using this method.

We invite the reader to observe the distribution of points, where the spatial proximity reflects the semantic similarity between the given publication and the authors' works. This method offers an intuitive way to understand the knowledge structure and the implicit connections between different research articles. We intend to incorporate all the above-mentioned changes into Label Studio to de-

crease the workload of the annotators and to increase the overall efficiency and accuracy of the process. The authors will be selected through the process of ranking and choosing the one with the highest similarity score for annotating the document. The author will be given a union set of 5 categories with respect to the author’s publications and the publication itself to choose from, which align the most with the publication.

F LLM Annotation

F.1 LLM Selection

In our work, we use the weights of the v1.1 model (https://github.com/lm-sys/FastChat/blob/main/docs/vicuna_weights_version.md) for Vicuna-7B and 13B. All inference tasks are executed on two sets of RTX 3090 (24GB of memory each). The parameters employed during response generation are as follows: `max_length` set to 100,000, `do_sample` enabled as True, and temperature adjusted to 0.7.

Our vision is to build a completely open source pipeline, so we have disregarded LLMs such as GPT-4 (OpenAI, 2023) which only provide API access, instead we have preferred open source alternatives such as LLaMA. We have explored non-LLaMA based LLMs such as OpenChatKit (TogetherComputer, 2023), but we encounter issues related to the stability of their output. We notice that these models sometimes produce inconsistent responses for the same data point (i.e., annotating one publication with keywords and labels), alternating between "Agree" and "Disagree" without providing logically coherent reasoning.

Among the multitude of LLaMA-based LLMs, we identify Vicuna as a model specifically fine-tuned for Question-Answering tasks, making it an apt choice for our project. Furthermore, Vicuna’s exceptional performance, underscored by its highest Elo rating in the Chatbot Arena (Zheng et al., 2023), convinces us to choose it as our annotator.

F.2 Single-Label

Figure 9 presents the standardized prompt template we employ to query the LLM regarding its agreement with the category predicted by our model for each data point in the ECON *single-label* dataset. It should be noted that certain segments of the prompts remain fixed and repetitive, a feature we refer to as "instructions". This design is necessitated by the LLM’s inherent propensity to forget

previous text, meaning that inputting the instruction just once may compromise the quality of responses for subsequent data points. For instance, they might cease to incorporate crucial keywords such as "Agree", "Disagree", "NOT ECON". Therefore, we find it essential to provide an instruction for each data point.

F.3 Multi-Label

For the ECON *multi-label* dataset, we engage the LLM with up to five prompts for each data point (Figure 10). The first prompt asks whether the given abstract is relevant to the field of Economics. If the answer is negative, we terminate further inquiry. However, if the LLM confirms the economic relevance, we proceed to query the model’s agreement with up to three categories our model had previously predicted. Lastly, we ask the LLM about any additional categories to which it believes the abstract may belong, beyond those predicted by our model. This final query is intended primarily as a preparatory measure for future keyword analysis. Similarly, the persistent recurrence of identical instructions within the prompt template is designed to mitigate the issue of forgetfulness inherent in the LLM. When we compare the output of the LLM with those of the human annotators on 100 annotations, we see a significant overlap in the categories allotted using both methodologies, which shows the potential to employ LLMs for multi-label tasks.

science-clf API **v1.0** **GA33**
 OpenAPI JSON
 API server for making inference on scientific text

default

POST /inference_by_model Model Inference

ask the api to predict text label

Parameters Try it out

No parameters

Request body **required** application/json

Example Value: Schema

```
{
  "text": "This is mathematics text.",
  "model_name": "model-1"
}
```

Responses

Curl

```
curl -X 'POST' \
  'http://localhost:8000/inference_by_model' \
  -H 'accept: application/json' \
  -H 'content-type: application/json' \
  -d '{
    "text": "This is mathematics text.",
    "model_name": "model-1"
  }'
```

Request URL
 http://localhost:8000/inference_by_model

Server response

| Code | Details |
|------|---|
| 200 | <p>Response body</p> <pre>{ "text": "This is mathematics text.", "prediction": [] }</pre> <p>Response headers</p> <pre>content-length: 38 content-type: application/json date: Wed, 22 Feb 2023 13:19:19 GMT server: uvicorn</pre> |

(1) API: inference_by_model.

science-clf API - Swagger UI

localhost:8000/docs#/default/batch_model_inference_batch_inference_by_model_post

POST /batch_inference_by_model Batch Model Inference

ask the api to predict text labels in batch

Parameters Cancel Reset

No parameters

Request body **required** application/json

```
{
  "text": "This is mathematics text.",
  "model_name": "model-1",
  "text": "This is a text snippet on public economics.",
  "model_name": "model-1"
}
```

Execute Clear

Responses

Curl

```
curl -X 'POST' \
  'http://localhost:8000/batch_inference_by_model' \
  -H 'accept: application/json' \
  -H 'content-type: application/json' \
  -d '{
    "text": "This is mathematics text.",
    "model_name": "model-1",
    "text": "This is a text snippet on public economics.",
    "model_name": "model-1"
  }'
```

Request URL
 http://localhost:8000/batch_inference_by_model

Server response

| Code | Details |
|------|--|
| 200 | <p>Response body</p> <pre>{ "text": "This is mathematics text.", "prediction": [13] }, "text": "This is a text snippet on public economics.", "prediction": [15] }</pre> <p>Response headers</p> <pre>content-length: 129 content-type: application/json date: Wed, 22 Feb 2023 13:18:13 GMT server: uvicorn</pre> |

(2) API: batch_inference_by_model.

Figure 7: Inference API Calls.

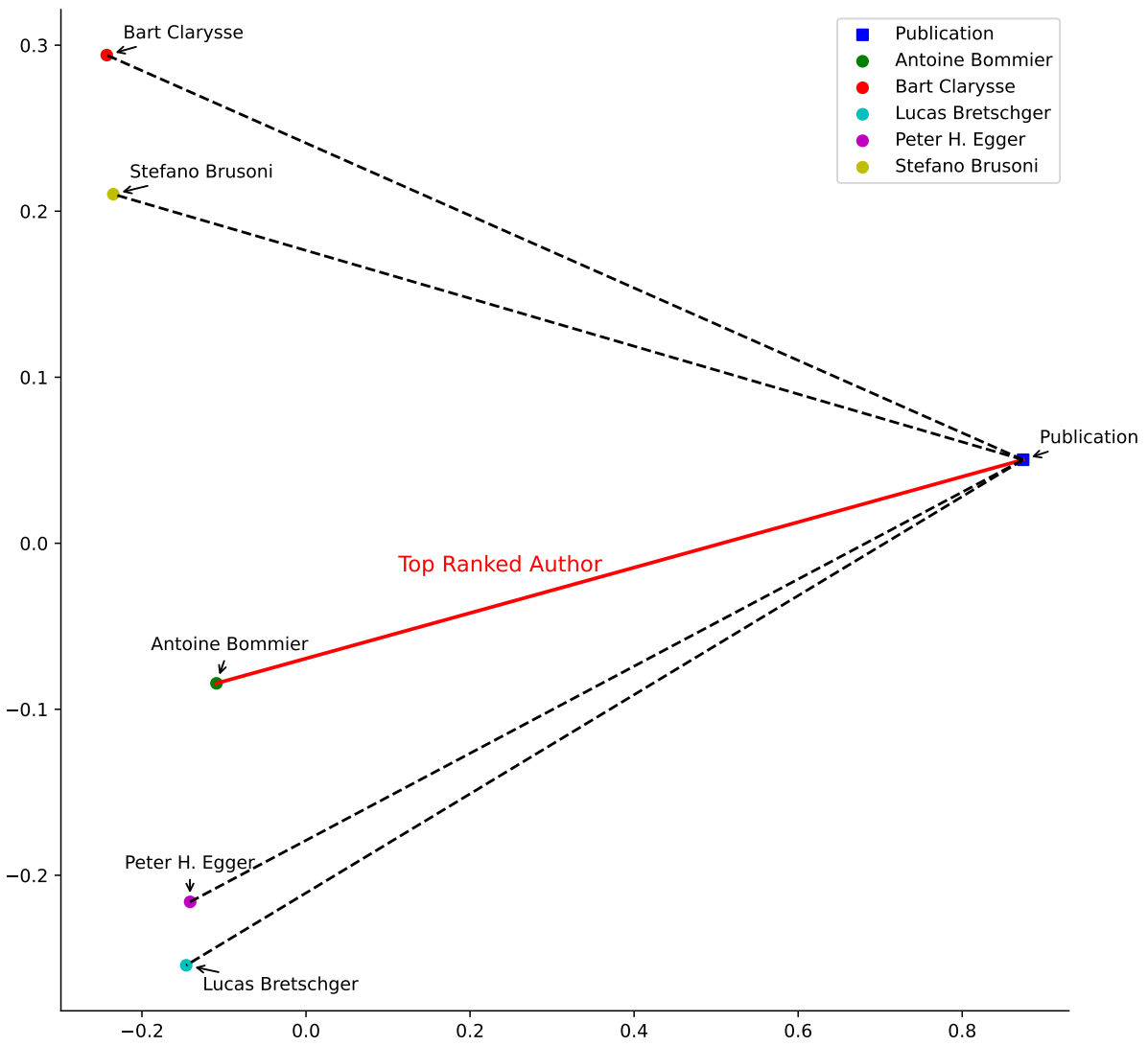


Figure 8: A Two-Dimensional Representation of the Semantic Proximity (between a specific publication and the top 25 cited author publications). Each point represents an author's work, and the spatial distribution reflects the semantic similarity to the given publication.

Prompt: I have trained a machine learning llm whose input is the abstract of a scientific article, and the predicted output is its predicted category.
Candidate categories include: {CAT1_NAMES}.
The abstract is: {abstract}.
My llm-predicted category is: {cat1}.
You have three options: 'Agree, Disagree, NOT ECON'.
Please choose ONLY ONE to output.
If you think this article belongs to the field of economics and the category predicted by the llm is correct, please output 'Agree' and give reasons;
Otherwise, If you think this article belongs to the field of economics but the category predicted by the llm is incorrect, please output 'Disagree', state which category it should belong to according to your opinion, and then provide reasons;
Otherwise, If you think this article does not belong to the field of economics, output 'NOT ECON', and give reasons.

Response: Disagree, Business Administration and Business Economics; Marketing; Accounting; Personnel Economics.
While the article does discuss adsorption/desorption behavior in alluvial soils, it also involves other economic factors such as cation exchange capacity and clay content, as well as the amount of Se desorbed in different forms. These factors suggest that the article may be more relevant to the field of Business Administration and Business Economics, which encompasses topics such as marketing, accounting, and personnel economics. Additionally, the article does not seem to fit squarely into the category of Agricultural and Natural Resource Economics & Environmental and Ecological Economics, as it is not solely focused on environmental or natural resource issues.

Figure 9: LLM Annotation Template for Single-Label Data. {CAT1_NAMES} should be replaced with a string list [Agricultural and Natural Resource Economics; Environmental and Ecological Economics, Industrial Organization, ...], which contains 18 elements (i.e., candidate categories). Taking Paper ID 2279482037 from the ECON single label dataset as an example, replace {abstract} with *Laboratory studies were conducted to investigate the adsorption/desorption behaviour ...* and replace {cat1} with *Agricultural and Natural Resource Economics; Environmental and Ecological Economics*. The response is the Vicuna-13B output for this data point.

1st Prompt: I have trained a machine learning llm whose input is the abstract of a scientific article, and the predicted output is its predicted categories (up to 3).

Candidate categories include: {CAT1_NAMES}.

The abstract is: {abstract}.

Do you think this abstract belongs to the field of Economics?

If so, output 'Yes'; if not, output 'No'.

1st Response: Yes.

2nd Prompt: The abstract is: {abstract}

Do you think this abstract belongs to the first model-predicted category {cat1}? If you agree with the first model-predicted category, please output 'Agree' and the reason; if you do not agree with the first model-predicted category, please output 'Disagree' and the reason.

2nd Response: Agree. The abstract describes a theoretical model of a monopoly firm that is introducing a new product with an unknown quality level to consumers. The model predicts that the firm will use price as a signal of quality, and that the price will monotonically increase with the quality level. This model falls within the category of microeconomics, which is concerned with the behavior of firms and consumers in markets, and how government policies can influence that behavior.

3rd Prompt: The abstract is: {abstract}

Do you think this abstract belongs to the second model-predicted category {cat2}? If you agree with the second model-predicted category, please output 'Agree' and the reason; if you do not agree with the second model-predicted category, please output 'Disagree' and the reason.

3rd Response: Disagree. The abstract does not appear to belong to the category of "History of Economic Thought, Methodology, and Heterodox Approaches." It is more closely related to the category of "Consumer Behavior and Demand Analysis" or "Market Structure and Pricing."

4th Prompt: Are there any other categories that you think are more suitable for this abstract, besides {cat1} {cat2} {cat3}?

If so, please output some other categories among candidate categories {CAT1_NAMES} and the reasons. If not, please output the reason why not.

4th Response: Some other categories that might be suitable for this abstract could include:

Agricultural and Natural Resource Economics: This category could be a good fit because the abstract discusses the impact of technology on the agricultural sector, including the use of drones for crop monitoring and automated harvesting.

Environmental and Ecological Economics: This category could also be a good fit because the abstract mentions the potential for drone technology to improve crop yield and reduce the use of pesticides, which could have positive environmental impacts.

However, it is also possible that the abstract could fit into multiple categories, depending on the specific focus of the research. For example, the abstract could be categorized as both Agricultural and Natural Resource Economics and Environmental and Ecological Economics. Overall, the categories of Business Administration and Business Economics, Mathematical and Quantitative Methods, Public Economics, and International Economics may not be as directly relevant to the focus of the research, although there could still be some overlap depending on the specific focus of the study.

Figure 10: LLM Annotation Template for Multi-Label Data. {CAT1_NAMES} should be replaced with the same string list described in Figure 9. Taking Paper ID 1497381249 from the ECON multilabel dataset as an example, replace {abstract} with *This paper analyzes the role of price as a signal of the quality of a monopoly firm's new product ...*, {cat1} with *Microeconomics*, and {cat2} with *History of Economic Thought, Methodology, and Heterodox Approaches*. Since our model only provides two categories, {cat3} is an empty string and no prompt is given. All responses are the outputs of Vicuna-13B for this data point.