

Toward Integrating ChatGPT Into Satellite Image Annotation Workflows: A Comparison of Label Quality and Costs of Human and Automated Annotators

Jacob Beck^{ID}, Lukas Malte Kemeter^{ID}, Konrad Dürrbeck^{ID}, Mohamed Hesham Ibrahim Abdalla^{ID}, and Frauke Kreuter^{ID}

Abstract—High-quality annotations are a critical success factor for machine learning (ML) applications. To achieve this, we have traditionally relied on human annotators, navigating the challenges of limited budgets and the varying task-specific expertise, costs, and availability. Since the emergence of large language models (LLMs), their popularity for generating automated annotations has grown, extending possibilities and complexity of designing an efficient annotation strategy. Increasingly, computer vision capabilities have been integrated into general-purpose LLMs like ChatGPT. This raises the question of how effectively LLMs can be used in satellite image annotation tasks and how they compare to traditional annotator types. This study presents a comprehensive investigation and comparison of various human and automated annotators for image classification. We evaluate the feasibility and economic competitiveness of using the ChatGPT4-V model for a complex land usage annotation task and compare it with alternative human annotators. A set of satellite images is annotated by a domain expert and 15 additional human and automated annotators, differing in expertise and costs. Our analysis examines the annotation quality loss between the expert and other annotators. This comparison is conducted through, first, descriptive analyzes, second, fitting linear probability models, and third, comparing F1-scores. Ultimately, we simulate annotation strategies where samples are split according to an automatically assigned certainty score. Routing low-certainty images to human annotators can cut total annotation costs by over 50% with minimal impact on label quality. We discuss implications regarding the economic competitiveness of annotation strategies, prompt engineering, and the task-specificity of expertise.

Received 6 August 2024; revised 30 September 2024 and 13 November 2024; accepted 4 January 2025. Date of publication 14 January 2025; date of current version 31 January 2025. The work was supported in part by the Munich Center for Machine Learning (MCML) and in part by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics — Data — Applications (ADA-Center) within the framework of BAYERN DIGITAL II under Grant 20-3410-2-9-8. (Jacob Beck and Lukas Malte Kemeter contributed equally to this work.) (Corresponding author: Jacob Beck.)

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

Jacob Beck and Frauke Kreuter are with the Munich Center for Machine Learning (MCML), Ludwig-Maximilians-Universität München Institut für Informatik, 80538 München, Germany (e-mail: jacob.beck@lmu.de; frauke.kreuter@lmu.de).

Lukas Malte Kemeter, Konrad Dürrbeck, and Mohamed Hesham Ibrahim Abdalla are with the Center for Applied Research on Supply Chain Services, Fraunhofer Institute for Integrated Circuits IIS, 90411 Nuremberg, Germany (e-mail: malte.kemeter@iis.fraunhofer.de; konrad.duerrbeck@iis.fraunhofer.de; mohamed.hesham.ibrahim.abdalla@iis.fraunhofer.de).

Digital Object Identifier 10.1109/JSTARS.2025.3528192

Index Terms—Automated annotations, ChatGPT, label quality, large language models (LLMs), satellite image annotation.

I. INTRODUCTION

WHEN collecting annotations for machine learning (ML) applications, researchers and practitioners must decide who will annotate the data. Ideally, annotations would already be available, or highly specialized experts could annotate sufficiently large datasets for model training, validation, and testing in-house. However, this is often not feasible for most applications, leading to paid (external) nonexpert annotators to handle the data annotation process. While methods from active learning or semisupervised learning (SSL) provide opportunities to minimize the required number of annotated samples when training models, a certain number of high quality annotations is still required and usually obtained from a human oracle. Furthermore, most real-world applications still require annotated data for validation and testing. Collecting high-quality annotations can be challenging even with a generous budget, as domain experts might simply not be available for large scale data annotation. This is especially true for many remote sensing use cases, where a high degree of task-specific expertise is required. High interclass similarity and intraclass variance are common issues and have arguably limited deep learning models in reaching their full potential in this domain [14]. These unique characteristics of satellite images are among the reasons why [51] consider remote sensing a “distinct modality” of ML. Given these conditions, annotators need to be experienced or well-trained in order to provide high-quality annotations. The scarcity and costliness of human domain experts calls for an exploration of the utility of other types of annotators. Generally, different types of annotators, ranging from domain experts to unskilled community volunteers, inherently vary in cost, availability, speed, and quality of their annotations. To the best of authors’ knowledge, there is no clear guidance on when to use which type of annotator or the implications for a project—especially now, with large language models (LLMs) entering the stage as an intriguing new option to consider for annotation.

The problem we address is clear: Given the need for annotated data and a limited annotation budget, whom should one pay to annotate satellite image data? One could prefer a domain expert, rely on laypersons, utilize a general purpose LLM like

ChatGPT—or even consider a mix. This decision can have strong implications for the resulting ML models. In this study, we examine and quantify the quality of various types of human and automated annotators. Each annotator classifies the same set of satellite images depicting industrial properties in Germany, identifying the usage status of each property. As land becomes increasingly scarce, it is in the public interest to minimize the use of new land consumption for industrial purposes and instead maximize the utilization of existing, underutilized, or abandoned industrial sites [10]. We refer to currently unused properties that can potentially be repurposed as *brownfields*. Using this practical example, our findings offer valuable insights and derive suggestions for incorporating LLMs into the annotation workflow.

Contribution: Given the wide range of available annotators, each with their own strengths and weaknesses, and the emerging trend of LLM annotations, the complexity of designing an effective annotation strategy has increased in recent years. Findings from previous research do not provide appropriate guidelines to navigate this decision problem. We address this research gap as follows:

- 1) assessing annotation quality and costs across various human and automated annotator types;
- 2) evaluating the feasibility and economic competitiveness of using the well-known ChatGPT4-V model for complex satellite image annotations;
- 3) highlighting the advantages of combining human and automated annotators through difficulty-based task allocation.

The rest of this article is organized as follows. Section II presents a discussion of related literature. The annotation tasks, data, and methods are introduced in Sections III and IV. Sections V and VI present results, which are discussed in Section VII. Finally, Section VIII concludes this article.

II. RELATED WORK

Complexity in image annotation: In this study, we consider a supervised image classification task for satellite images. While training image classification models is well studied [35], [37], [61], the preceding image annotation process is complex and under-researched. For many applications, it is not straightforward to define the annotation convention, which classes to consider and what granularity level to choose. A prominent example for the difficulty of image annotation is the ImageNet dataset, a classic benchmark for image classification tasks [15]. Its annotations have been shown to contain errors caused by unclear conventions, class ambiguity (e.g., “sunglass” versus “sunglasses”) and ignoring the possibility of multilabels (e.g., ant and ladybug on the same image) [6], [67]. Related research found 6% of the validation set to be mislabeled [43]. Especially in the context of satellite images, annotation can be difficult. Remote sensing data often comes with high intraclass variance and high interclass similarity, which is argued to make the common approach of using the majority vote from different nonexpert annotators less effective [62]. As another complicating factor, visual features for land coverage (LC) classification are often not sufficient for land usage (LU) classification [70]. Certain objects like trees or cars may be present across all LU classes. The absence or presence of certain things in an

image does not necessarily allow to infer LU. Instead, it is the interplay between various LC classes and the overall context and composition that determines the correct LU class. We have shown in previous work that crowd-sourced annotations from OpenStreetMap (OSM) [46] can be considered as a source for easily accessible LU annotations [20]. Using OSM data for ML annotations is not uncommon but comes with its own quality concerns [22], [25], [59].

Human annotations: When developing an annotation strategy, multiple different annotator types, ranging from experts to laypersons, are usually considered. This consideration is nontrivial since humans differ in their behavior of assigning annotations to a set of instances. Most fundamentally, two different annotators will hardly ever assign the exact same annotations to a dataset. Disparities in annotations have been shown to be correlated with a variety of annotator characteristics and annotation task setup, a phenomenon that can be described as *annotation sensitivity* [4], [13], [29]. That means simple manipulations like varying the order in which annotations are requested impact the resulting annotations [5] and the resulting models [29]. Another example is the high importance of annotation guidelines, which are commonly provided at the beginning of an annotation task [40], [41]. In addition, task-specific cognitive bias affects how an individual completes an annotation task. This can manifest itself for example in increased rates of nonresponse or “straightlining” (always selecting the same choice) and can reduce data quality [30]. These biases occurring in the annotation process are likely to manifest themselves differently for different types of human annotators. While laypersons might feel like they have to build their reasoning from scratch, experts can rely on their existing knowledge and beliefs [27]. A meaningful body of research has assessed the quality gap between expert and nonexpert annotators [3], [19], [44]. The results show that for many annotation tasks, nonexperts can come close to expert annotation quality when building on a variety of additional methods. Moreover, using the majority vote of multiple annotators can improve overall annotation quality [44]. Similar annotation quality levels across laypersons and experts could be found in other applications like occupation coding [38] or image annotation for organ segmentation [28]. Further studies have creatively investigated other possibilities to leverage layperson annotation in expert domains [60], [63]—e.g., by transforming an expert-level task (fish recognition task) to an easier one (visual similarity task) to allow for layperson annotation [26]. Despite previous research on image annotation tasks across various domains, there appears to be limited evidence and empirically validated best practices for collecting human annotations of satellite images.

Automated Annotations: As acquiring high-quality annotations with human annotators is costly, complex, and prone to errors, the desire for automated annotation techniques is high. Methods like confident learning (CL) [42] try to automatically identify and counteract noisy labels in already annotated data. In addition, methods from SSL can be applied to reduce the demand for labels in model training by automatically pseudolabeling the available unsupervised data. Such methods leverage the few available annotated samples to make predictions on the unlabeled pool [16], [36]. Usually, only pseudolabels with very confident predictions are kept for training the

model. Many state-of-the-art frameworks such as FixMatch [54], FlexMatch [68], and OpenMatch [52] rely on pseudolabeling. However, pseudolabeling is shown to suffer from confirmation bias [2]—especially when the data contains real-world challenges [21]. It alone will, therefore, not eliminate the need for human annotations. While our work is not focused on pseudolabeling directly, we investigate whether LLMs can serve as valid automated annotation tools. Since LLMs are foundation models pretrained on a large amount of data and not directly retrained on their own predictions, they inherit promising potential for data annotation. Since the emergence of LLMs such as the ChatGPT models, multiple studies have assessed their capabilities to provide annotations. Due to the delayed dissemination of vision-models, most research has tackled text-based tasks, mainly in the NLP area. Here, ChatGPT models have been shown to provide comparable annotation data quality [17], [65] or even perform better than human annotators [58]. At the same time, LLMs inherited strong biases or even responded in a random fashion in survey question answering tasks [18]. Furthermore, the quality of the resulting annotations seems to be strongly data- and task-specific [49].

LLMs and vision-language models for remote sensing: Recently, specific works for geo-spatial analysis in combination with LLMs have emerged. GeoGPT [69] uses an LLM to develop a framework for performing GIS operations based on language instructions. GPT4GEO [50] tests ChatGPT-4 on navigation and planning tasks. Li et al. [33] provided a good overview of vision-language modeling for remote sensing and applications like scene classification [34], [47], image captioning [71], and image retrieval [66].

In this study, we consider three aspects to be especially important to highlight. First, it can be advantageous to mix different annotator types. Related works in this context often compare human expert versus human layperson annotators (discussed above). In addition, first promising efforts have been made to augment human annotation capabilities through the addition of LLMs [23], [32] or to replace human annotators with LLMs [1], [58]. Second, we allocate annotation tasks to different annotators based on the estimated task difficulty. Other work has previously entertained the idea of predicting annotation task difficulty beforehand and routing the instance to a suitable annotator type [63]. Our proposed approach in Section VI was inspired by this concept. Third, we employ the self-reported certainty scores from automated annotation tools for this prediction of task difficulty. Especially for LLMs, the usefulness and reliability of self-assessments is unclear and an active research area [11], [12], [53], [64].

III. DATA

Recall that our main substantial objective is to identify the LU status of industrial areas. For Germany, information is available regarding which areas are designated for industrial usage, including corresponding property boundaries [7], [8]. Such a designation regulates legal land use and does not provide any information about whether the property is actually actively used for business. For this study, a subset of 2955 properties was randomly selected. For each property, a corresponding satellite image is acquired from different image providers: Google [24],

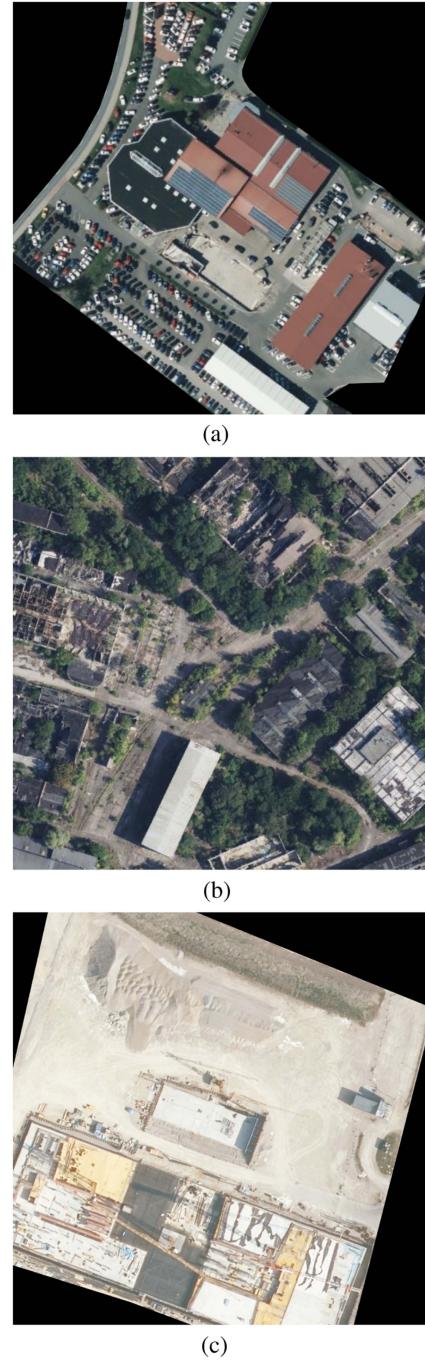


Fig. 1. Illustration of unambiguous examples for the three classes. To address high intraclass variance and interclass similarity, the annotation guidelines (see Fig. 10 in the Appendix) also incorporate more details on potential variety within the classes. Source and license: GeoBasisDaten/BKG (2024). (a) Active. (b) Brownfield. (c) Construction.

Bing [39], and official digital orthophotos with a ground resolution of 20 cm [9] (for details see Fig. 9 in the Appendix).

Annotation task: Annotating our data requires an accurate matching of the visual features of a property (image) and its usage status (label). The annotation task is to classify an image's LU status as either *actively used land* (short: *active*), *construction*, or *brownfield* (for example, see Fig. 1), in accordance with the official guidelines of the State Agency for Nature, Environment, and Consumer Protection North Rhine-Westphalia

(LANUV) [31]. The officially recommended method for identifying *brownfields* also includes in-person visits and the use of ownership databases or other meta-data. As these additional sources of information are not available, annotators had to rely solely on the visual information from the images, making this task fairly difficult. For gold-standard annotations, a single domain expert from the company *Spacedatists GmbH* annotated all data. This company specializes in identifying brownfields on satellite images and has conducted similar tasks on behalf of federal state governments in Germany before. We refer to the annotations as “expert annotations.” The experts additionally provide a certainty score (from 1 = “very uncertain” to 4 = “very certain”).

Data preprocessing: A rectangular aerial image of a property might include parts of neighboring properties, but the usage status of the property in question should be independent of its neighbors. Displaying the entire image to annotators can lead to confusion, as the assigned annotation would only be accurate for a portion of the image. To address this issue, properties are cropped and neighboring properties are masked with a black background using publicly available boundary data. This ensures that the displayed property corresponds accurately to the assigned label, following the methodology outlined by [20]. Since each image corresponds to a specific land parcel, we assume that only one LU status can be present at a time for each parcel. For very large properties that exceed the defined grid size, only the center of the property is displayed.

Human annotation: We collect nonexpert human annotations from ML engineers with medium domain knowledge,¹ trained research assistants with low domain knowledge, and laypersons with no domain knowledge (three individuals for each annotator type). The three groups of human annotators differ with respect to domain knowledge, methodical knowledge, and costs. To aggregate the three labels per annotator type, we compute a majority vote—a standard procedure in annotation for model training purposes. All nonexpert annotators received the same class definitions and decision rules as the expert (see Fig. 10 in the Appendix). The annotators completed the task in LabelStudio (licensed under Apache-2.0) [57] and had the opportunity to ask questions regarding both the classes and the tool before the start of the study.

Automated Annotation: We use three additional types of automated annotations. First, we extract information about each property’s land use status from OSM. OSM provides open data under an *ODbl* license. The annotation process follows a workflow proposed in previous work [20]. Properties are annotated automatically by querying the OverpassTurbo [48] API for the tag “*landuse*” and mapped to the corresponding satellite images by our pipeline.² We refer to these as *OSM pipeline annotations*, though they are, in fact, voluntarily provided, crowd-sourced labels.

¹ They gained experience in developing classification models for satellite image data from past projects; however, they are not domain experts.

² Note that on OSM, *brownfield* and *construction* are common tags that members of the community assign. However, not every property gets tagged and especially actively used properties do not always get tagged as such. Hence, our labels for the *active* class were additionally validated by checking if any company is listed for the address in public company registers [20].

TABLE I
SUMMARY OF COSTS PER ANNOTATION FOR DIFFERENT ANNOTATOR TYPES (IN GERMANY)

Annotator type	Abbreviation	Approximate cost per annotation
Domain expert	Expert	\$0.55
ML engineer	ML	\$0.27
Research assistant	RA	\$0.07
Layperson	LP	\$0.02
ChatGPT4-V	GPT	\$0.01
OSM pipeline	OSM	\$0.00
Confident learning	CL	\$0.00

Converted from EUR to USD.

As a second source of automated annotations, we utilize the well-established method of CL, which addresses noisy annotations for a given annotated dataset by either pruning incorrect labels or proposing corrections based on uncertainty estimates [42]. It is an appealing approach to cure issues in erroneous annotations with CL to acquire improved annotations. The method also comes with a well-maintained GitHub repository [56] under a *AGPL-3.0-or-later* license. CL is designed to detect and clean label issues in provided data. As it requires a vector of annotations as input, we use the noisy OSM annotations. The method identifies label issues and provides proposals for correction. We refer to the proposed cleaned annotations as *CL annotations*. By default, CL outputs a certainty (or label quality) score between 0 and 1 for each label. For a detailed description of the annotation workflow in CL, we refer to the original paper [42]. Finally, we employ the OpenAI ChatGPT4-V API to collect automated annotations from a general-purpose LLM [45]. Here, we route the same set of images together with a prompt to the API and collect a standardized response. In this process, we experiment with low (0.2) and high (0.7) temperature settings and by using both a short, naive prompt and a more elaborate, longer prompt. The long prompt included class definitions similar to those provided to the nonexpert human annotators and decision rules. Both prompts requested a short reasoning and a certainty score from 1 to 4 (on the same scale as the human expert). It remains unclear whether an LLM can accurately judge its own certainty or provide reliable reasoning for its decisions. These additions are considered exploratory elements in experimenting with this novel method. The exact prompts can be found in Figs. 11 and 12 in the Appendix.

Economic considerations: We estimate and report annotation costs per instance to assess the economic competitiveness of different annotator candidates. Table I provides an overview of all considered human and automated annotators, along with their approximate costs per annotation.

Note that we only consider the marginal costs of generating an additional annotation and implementation and computational costs are ignored. We denote marginal costs as $C_{\text{annotator}}$. Development or set-up costs for the pipeline, the CL implementation, or the ChatGPT4-V API script are disregarded. In light of ethical considerations and the working conditions of human annotators, note that the crowd-sourced OSM labels are assigned by volunteers, and all research assistants and laypersons were formally employed under German law.

IV. METHODS

The data collection process resulted in 16 sets of annotations for the 2955 images: one from an expert, three different human annotator types with three individuals each, four different GPT configurations, one from CL and one from OSM. Annotating data by paying a domain expert is considered the default annotation strategy. Our setup now allows for an investigation of the cost-saving potential of replacing the expert with alternative annotator types. Given the annotation cost and quality interplay, we investigate different aspects of label quality in this study. Most importantly, we report and evaluate the agreement of our candidate annotators with the domain expert. We further assess, which factors influence and explain agreement. Eventually, we use our data for training CNN models to evaluate downstream task performance of the various annotators.

A. Assessment of Annotator Agreement

We begin our analysis with a descriptive investigation of agreement rates. Agreement on our three-class problem checks if the votes v_1 and v_2 of two annotators are equal. Equation (1) defines image-level agreement, we later report aggregated agreement rates across all images in our set

$$\text{For } v_1, v_2 \in \{0, 1, 2\}, \quad A(v_1, v_2) = \begin{cases} 1, & \text{if } v_1 = v_2 \\ 0, & \text{if } v_1 \neq v_2 \end{cases} \quad (1)$$

with 0, 1, 2 referring to the three LU classes. Subsequently, we look at agreement between the expert and our candidate annotators grouped by certain relevant features such as expert certainty or expert label.

B. Statistical Analysis

Second, we fit linear probability models for each of the distinct annotator types. A binary indicator for agreement with the expert annotation (0 = No, 1 = Yes) serves as the dependent variable. Regression coefficients can, thus, be interpreted as changes in probability of agreement with the expert. The set of explanatory variables varies between annotator types. Variables included in all of the models are image source, expert label, annotator label, and expert certainty.³

C. Training of DNNs

In addition to the regression models, we approximate annotator quality by training deep learning models on annotations from different annotator types. Recall that the 2955 images described in Section III were annotated by all annotator types as well as the expert. In Section V-C, these data are split into five folds, with four folds used for training and one for testing. We train 7 CNN models with the same parameters and 5-fold cross validation (CV) under full supervision. Each model uses the labels from a different annotator type for training. The expert annotations for the fifth fold are used as the ground truth for all combinations. The resulting test scores are then aggregated across the five different CV runs to obtain the reported F1-score. We report the corresponding CNN training parameters in Table VI in the

³ Additional information on the full set of regression parameters can be found in the Appendix.

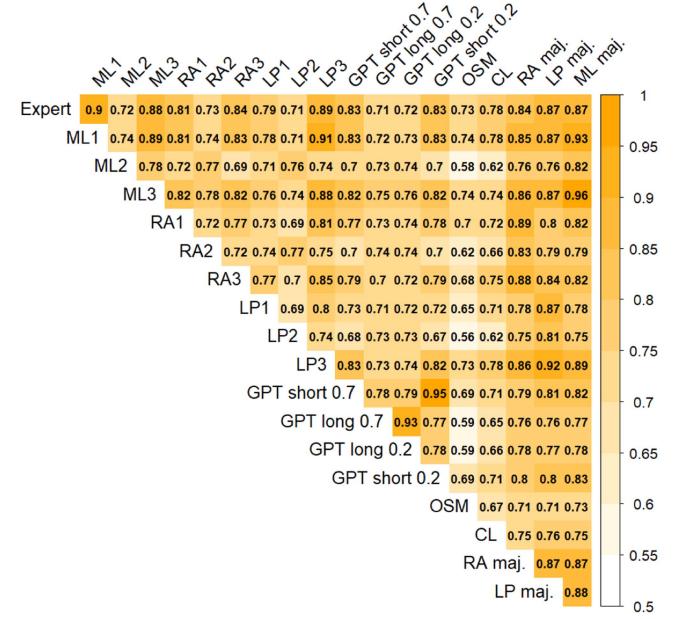


Fig. 2. Agreement matrix for all annotator types.

Appendix and F1-scores for *brownfields* on expert-annotated test sets in Table III.

D. Comparing Annotation Strategies

In Section VI, we compare various annotation strategies. Previously, we focused on comparing individual annotator types. In this section, we shift our attention to evaluating whether combining different annotator types in a hybrid approach can be a beneficial strategy. We propose a routine to calibrate automatically generated labels (from LLMs or CL). By leveraging their self-reported certainty scores, ambiguous instances can be rechecked by a human annotator. We will provide the algorithms and formulas for calculating costs and expert agreement rates (quality) associated with each strategy.

V. RESULTS

This section is divided into the following three sections. We begin by comparing the agreement scores among different annotator types in Section V-A. This is followed by an in-depth analysis of influencing factors in Section V-B. Section V-C reports the test scores for models trained on annotations from each annotator type to assess how annotation quality impacts a downstream prediction task.

A. Agreement Scores

The expert classified 2020 (68%) images as *active*, 656 (22%) as *brownfields*, and 279 (~10%) as *constructions*. We compare agreement scores of all other individual annotators with the expert in row 1 of the full agreement matrix displayed in Fig. 2. The highest agreement rate A is observed for ML1 and, surprisingly, LP3. These two best individual annotators are immediately followed by the majority votes of the ML engineers (ML maj.) (87%), laypersons (LP maj.) (87%), and research

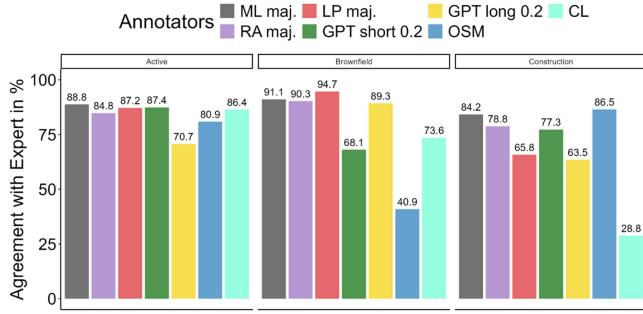


Fig. 3. Agreement by expert label. Overall agreement highest for *active*, human annotators detect *brownfields* particularly well.

assistants (RA maj.) (84%). We observe low agreement of 73% for the automated annotation from the OSM pipeline. The CL approach is able to account for some of the noise in the OSM labels, increasing the agreement to 78%. Finally, the four GPT configurations varying in temperature settings (0.2 versus 0.7) and the sophistication of the prompts (short versus long) achieve an agreement ranging between 71% and 83%. Interestingly, the temperature had no effect on the agreement scores, while “naive” prompting clearly resulted in higher expert agreement compared to longer, more complex prompts.

B. Deep Dive Into Influencing Factors

To investigate further, Fig. 3 presents the agreement between annotator types and the expert, categorized by the assigned expert label. Generally, we observe the highest agreement for the *active* class, with all agreement rates exceeding 70%. Human annotators⁴ showed the least agreement for the *construction* class while achieving the highest agreement for *brownfields*. For the automated annotators, the results vary: the agreement between the expert and the OSM annotations is very low for *brownfields*, whereas the CL approach scores particularly low for *constructions*.

An interesting pattern emerges for the GPT annotations. While the short prompt (dark green) showed better overall agreement rates, Fig. 3 reveals that this is mainly due to correct predictions in the *active* and *construction* classes. The long-prompted GPT annotator (yellow) had a lower overall agreement but performed nearly as well as the human annotators in identifying *brownfields*. The naive prompting performed poorly in the *brownfield* class, but since *active* occurs more often, its overall agreement is higher. Due to this imbalance, an annotator could achieve 68% agreement by categorising all instances as *active*. Thus, the sophisticated prompt enhanced the LLM’s ability to detect *brownfields* at the expense of agreement for the other classes. Overall, the three classes vary in difficulty, with human and automated annotators demonstrating varying levels of proficiency in identifying different classes. However, agreement might not only differ across classes but also based on instance-level difficulty. We use the expert’s self-reported confidence score as a proxy for difficulty and assess its relationship to expert agreement in Fig. 4. Here, a clear trend can be observed: as expert certainty increases, agreement improves across all annotator types. This indicates that the

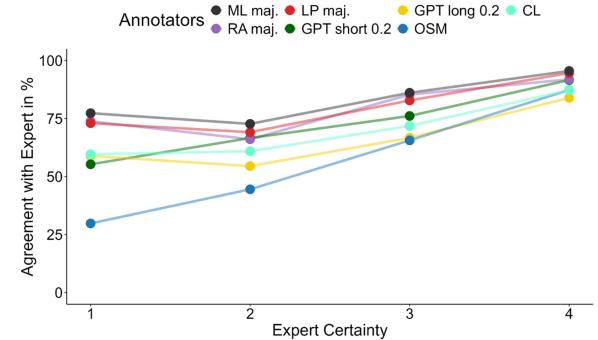


Fig. 4. Agreement rates by expert certainty. With increasing expert certainty, agreement generally increases.

certainty score is an effective measure of instance-level task difficulty.

Especially the crowdsourced OSM labels show a notable increase in agreement as expert certainty rises suggesting that crowdworkers perform well for easier tasks and instances but may struggle with more difficult ones. Given that the images originated from various sources, we examined whether there are any source-specific differences in agreement. Our analysis revealed no noteworthy discrepancies (see Fig. 13 in the Appendix). Table II summarizes all regression coefficients of the annotator type-level linear probability models. The analysis reveals that, across all annotator types (except for CL), the probability of agreeing with the expert is significantly lower when annotators assigned the *brownfield* or *construction* classes compared to the *active* class. Notably, the magnitude of these coefficients is larger for the *construction* class. For automated annotators like GPT and OSM, the probability of agreement with the expert decreased on average by 30% to 60% when annotating instances as *construction*. This finding is somewhat expected, given the true class distribution where most images are categorized as *active* by the expert. However, even after controlling for the assigned expert annotation, it is clear that annotators are generally less likely to be correct when labeling instances as *brownfield* or *construction*. As Fig. 4 previously suggested, increasing expert certainty is significantly related to a higher probability of agreement. Images where the expert is very certain appeared easier for other annotator types to label accurately. The image source had no impact on agreement, which is a desirable outcome, as it indicates that a model trained on these annotations would not overfit to any specific image source. These results align with previously presented findings but are more robust as the regression model controls for important covariates. Annotation time and order are largely unrelated to expert agreement in the models where they are included. Interestingly, the certainty score provided by GPT is strongly and significantly associated with an increased probability of agreement. This finding emphasizes that the certainty assessment by GPT is likely to be a valuable indicator of the correctness and uncertainty of a label. The same holds true for the CL label quality score, which yields a large and significant regression coefficient. Subsequently, we examine the usefulness of certainty scores in more detail in Section VI.

⁴ For better readability, we report majority votes for human annotator types and ChatGPT results with a temperature of 0.2.

TABLE II
REGRESSION COEFFICIENTS WITH SIGNIFICANCE LEVELS

Coefficient	Models						
	ML	RA	LP	GPT short	GPT long	OSM	CL
Annotator Label: BR	-0.13***	-0.17***	-0.29***	-0.14***	-0.21***	-0.18***	-0.03
Annotator Label: CO	-0.12**	-0.21***	-0.18***	-0.40***	-0.30***	-0.60***	-0.03
Expert Label: BR	0.00	-0.18***	-0.02	-0.07**	-0.06***	-0.11***	0.03
Expert Label: CO	0.02	-0.10*	-0.09*	0.05	-0.05	0.06*	0.04
Expert Certainty	0.05***	0.03***	0.02**	0.06***	-0.00	0.09***	0.02***
Source: DOP	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Source: Google	-0.01	-0.00	0.00	0.00	0.00	0.01*	0.00
Median Annotation Time	-0.02***	0.00	-0.00				
Mean Annotation Order		-0.00	-0.00				
GPT Certainty				0.08***	0.22***		
CL Label Quality							0.82***
(Intercept)	0.82***	0.89***	0.93***	0.44***	0.07	0.62***	0.08**

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Coefficient magnitude can be interpreted as probability change to agree with the expert. For GPT the temperature is 0.2.

TABLE III
5-FOLD CROSS-VALIDATED PERFORMANCE

Test CV F1-Score	OSM	CL	GPT short	GPT long	LP maj.	RA maj.	ML maj.
for Brownfield class	0.37	0.68	0.61	0.75	0.75	0.71	0.77
for all classes	0.31	0.58	0.66	0.69	0.68	0.66	0.71

Columns refer to the source of training data annotations. The GPT temperature is 0.2. For each CV iteration, 4 folds are used for training and one for testing on the expert labels.

C. Assessing Impact on Model Performance

The overall objective of any annotation task is to train high-quality ML models with the annotated data. So far, we have only considered agreement with an expert as a proxy for label quality. To directly assess the effect of label quality on model performance, we perform 5-fold CV on our data as described in Section IV. The goal of this analysis is to evaluate how well training annotations from each annotator type perform in correctly predicting *brownfields*. We report F1-scores of individual models trained on 4 folds with the labels provided by one annotator candidate in Table III. For each CV loop, the model is tested on the expert annotation for the fifth fold. The table reports the aggregated F1-scores from all 5 CV runs. The results align with our previous findings: the noise in the OSM labels adversely affects the F1-score, human annotators perform at similar levels, and a well-designed ChatGPT prompt can compete with human annotators.

VI. TOWARD INTEGRATING CHATGPT IN THE ANNOTATION WORKFLOW

We have now thoroughly compared different types of potential annotators for a satellite image annotation task against each other. Recall that previous studies have demonstrated potential in routing samples to different annotators based on instance-level difficulty. In our study, the expert, the LLM, and CL all provide some version of a confidence score. Our results in Table II have shown that such scores can be considered as a good proxy measure for instance-level difficulty. In this section, we propose a routine to cost-efficiently integrate ChatGPT (or any similar LLM) into the annotation workflow. The idea here is to first get

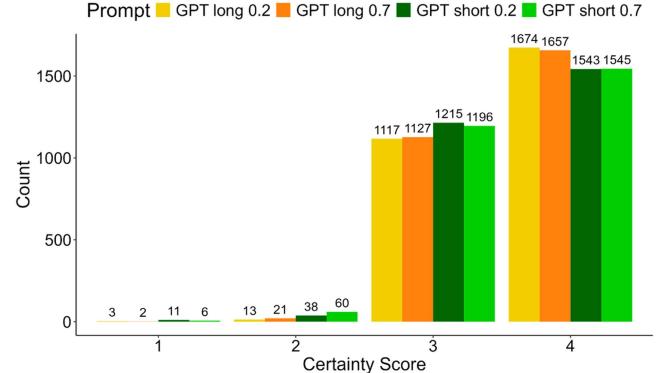


Fig. 5. GPT certainty distribution. All four GPT annotators almost always assigned a certainty score of 3 or 4.

labels for the whole set of images from a cheap automated annotator, assess the difficulty of an instance through the certainty score and then route ambiguous instances to a human annotator for reannotation.

First, consider descriptive analyses of the self-reported certainty scores. The certainty score distribution of ChatGPT in Fig. 5 shows that scores of 1 or 2 occurred very rarely, with almost all annotations assigned a score of 3 or 4. In terms of expert annotator agreement, Fig. 6 demonstrates a significant increase in agreement between ChatGPT certainty scores of 3 and 4 across both temperature modifications and prompt types. Most notably, when ChatGPT assigned the highest certainty score (which occurred in more than half of the instances), its label agreed with the expert label in over 90% of the cases across all four modifications.

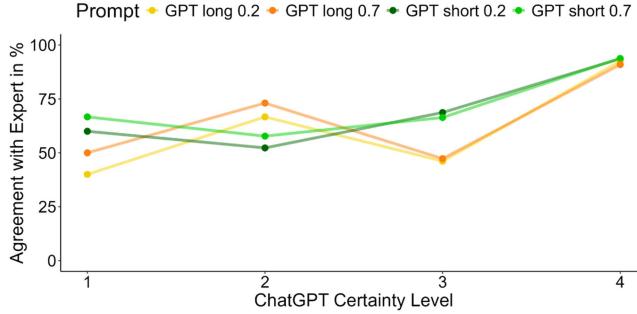


Fig. 6. Agreement rates by ChatGPT certainty.

The question is now, if we can use the certainty scores for efficiently combining cheap automated annotators with expensive human annotators. Our study setup allows for a comparison of dollar-value costs and agreement rates for various simulated annotation strategies. For notations, we use A to represent agreement with ground truth labels, S_i for annotation strategy i , X for the number of samples, and C as costs. We will compare the following four simulated strategies.

- 1) S_1 : All samples are annotated by the expert.
- 2) S_2 : All samples are initially annotated by ChatGPT. Samples with a ChatGPT-certainty score below 4 (very certain) are re-annotated by the expert.
- 3) S_3 : All samples are initially annotated by ChatGPT. Samples with a ChatGPT-certainty score below 4 (very certain) are re-annotated by a nonexpert human annotator. Only when ChatGPT and the nonexpert disagree are the samples passed to the expert for final annotation (see Algorithm 1).
- 4) S_4 : All samples are annotated solely by ChatGPT.

In addition, we consider $\theta(x)$ to be the self-reported ChatGPT certainty score for sample x and $L(x)$ to be the label of sample x .

Algorithm 1 outlines the proposed hybrid strategy S_3 . The algorithm can also be used to deduct S_2 by reducing the *else* clause in lines 6–12 according to Algorithm 2 in the Appendix. The difference is that for S_2 , all images with a certainty score $\theta < 4$ will be routed directly to the expert. The four strategies can now be compared with respect to agreement rate A and the corresponding cost C . Since the expert annotations are considered to be our ground truth label, the baseline strategy S_1 corresponds to perfect agreement of 1 at a cost of 1625 USD. Expert agreement for strategy S_4 can be inferred from Fig. 2, and total costs can be calculated by multiplying the total number of images X with the marginal costs C_{GPT} listed in Table I. For S_2 and S_3 , consider (2) to (3)

$$C_{S_2} = |X| \cdot C_{\text{GPT}} + |\{x \in X \mid \theta(x) < 4\}| \cdot C_{\text{expert}} \quad (2)$$

$$C_{S_3} = |X| \cdot C_{\text{GPT}} + |\{x \in X \mid \theta(x) < 4\}| \cdot C_{\text{non_expert}} + |\{x \in X \mid \theta(x) < 4 \wedge L_{\text{GPT}}(x) \neq L_{\text{non_expert}}(x)\}| \cdot C_{\text{exp}} \quad (3)$$

TABLE IV
COST AND AGREEMENT COMPARISON FOR HYBRID LABEL CALIBRATION WITH CHATGPT VARIANTS

Prompt	Temp.		S_2	S_3	S_4
long	0.2	Estimated cost	697	442	29
		Agreement	0.95	0.88	0.72
long	0.7	Estimated cost	705	444	29
		Agreement	0.93	0.87	0.71
short	0.2	Estimated cost	772	390	29
		Agreement	0.96	0.91	0.83
short	0.7	Estimated cost	772	397	29
		Agreement	0.96	0.91	0.83

TABLE V
ALTERNATIVE HYBRID CALIBRATION CL INSTEAD OF LLM LABELS

CL Ablation	S_2^{CL}	S_3^{CL}	S_4^{CL}
	675	416	0
	0.97	0.93	0.78

Achieved by replacing $L_{\text{GPT}}(x)$ in Algorithm 1 With $L_{\text{CL}}(x)$ and applying a certainty score cutoff of > 0.99 (see updated Algorithm 4 in the Appendix).

Algorithm 1: Hybrid Strategy S_3 .

```

1: Start with a pool of unlabeled samples  $X$ .
2: for each sample  $x$  in  $X$  do
3:   compute  $L_{\text{GPT}}(x)$  and  $\theta(x)$ 
4:   if certainty score  $\theta(x) == 4$  then
5:      $L_{S_3}(x) = L_{\text{GPT}}(x)$ 
6:   else
7:     compute  $L_{\text{non\_expert}}(x)$ 
8:     if  $L_{\text{GPT}}(x) == L_{\text{non\_expert}}(x)$  then
9:        $L_{S_3}(x) = L_{\text{GPT}}(x)$ 
10:    else
11:      compute expert label  $L_{\text{expert}}(x)$ 
12:      set  $L_{S_3}(x) = L_{\text{expert}}(x)$ 
13:    end if
14:  end if
15: end for

```

$$A_{S_i} = \frac{|\{x \in X \mid L_{S_i}(x) = L_{\text{expert}}(x)\}|}{|X|}. \quad (4)$$

We consider all four different variants of ChatGPT for calibration with varying prompts and temperature. For the nonexpert human annotator, we will always use “Research Assistant 1” as the middle performer in that annotator group. The results shown in Table IV underline how beneficial and cost-saving the integration of a general-purpose LLM annotator would have been for our image annotation workflow in a counterfactual scenario.

Across all four modifications of ChatGPT, we are able to drastically reduce annotation costs with a relatively small decrease in label quality. For strategy S_2 , agreement with the experts remains well above 90%, but total costs are reduced by over 50%. Strategy S_3 offers higher agreement compared to S_4 , corresponding to a cost reduction of around 75% against S_1 while maintaining an overall agreement just below 90%. We

achieve this by bucketing our data based on difficulty inferred from the self-reported LLM certainty score. Easy images are labeled by the LLM, harder instances are relabeled a second time by one or more human annotators. These results strongly indicate that integrating LLM annotators in a hybrid manner can be a highly valuable approach for projects with strict budget constraints. Given that LLMs are not the only available source of automated annotations we could use for this approach, Table V reports an ablation with using CL labels instead of the ones provided by ChatGPT. As instance difficulty indicator we use the label quality score reported by CL. As the distribution of this score in our data appeared to be heavily skewed toward the tails 0 and 1 (see Fig. 14 in the Appendix) we categorize instances with a quality score ≥ 0.99 as certain. In these instances, the CL annotations matched the expert's in more than 95% of cases.

Interestingly, we find that deploying CL with a certainty cutoff of 0.99 outperforms our four LLM variants, presented in Table IV, for strategy S_2 . S_2^{CL} is cheaper and comes along higher agreement than its LLM competitors. In both cases, a similar number of samples is routed to humans for reannotation. However, for the images that are not reannotated, CL has a higher agreement than all four GPT configurations. Also recall that we assumed the marginal costs of CL to be 0.00, while retrieving ChatGPT annotations for all 2955 images costs around 29 USD in total—which partially explains the cost advantage of CL. For strategy S_3 , results are mixed. CL provides higher label quality at slightly higher costs compared to the short ChatGPT prompts, while outperforming the long prompts in costs and agreement rate. It is to be noted as an argument for LLMs that, in contrast to CL, they work in zero-shot fashion and no model training is needed. The simulation of these hybrid strategies suggests that our proposed routine of integrating LLMs or other automated annotators into the annotation workflow can decrease annotation costs with only minimal loss of label quality. Since the certainty score is a fundamental element of our proposed routine, it is worthwhile to investigate how ChatGPT judges specific images in more detail. Fig. 7 presents a few selected examples alongside the corresponding annotator choices to address three main questions. First, we display images with low LLM certainty scores in row 1 of the figure. Second, we show images that received the highest certainty but on which LLM and expert disagreed in row 2. This corresponds to the agreement gap between strategies S_1 and S_2 . Third, we consider images which would be passed to the expert following strategy S_3 and Algorithm 1 in row 3. For all instances, ChatGPT was instructed to provide a reasoning for its classification, which can be found in Table VII in the Appendix.

For the “GPT long 0.2” configuration, only 16 images received a certainty score of 1 (very uncertain) or 2 (uncertain). Upon visual inspection, these images tend to be small and with low color-diversity. Three examples are provided in Fig. 7.

Subsequently, we want to highlight a few observations based on the reasoning provided by ChatGPT. The first image in row 2 shows what can be considered a material yard. The property seems to be in active usage but contains features very similar to a construction site. ChatGPT correctly identifies these similarities (“Earth-moving activities on the left and scattered construction material”) but does not recognize the concept of a material yard as an active property. The expert classified it as *active*. The second image in this row is annotated as a *brownfield* by the expert,

but ChatGPT assigned the *construction* class due to the presence of an excavator (see reasoning in Table VII in the Appendix). It is unlikely that the expert missed the excavator, but they apparently considered the heterogeneous vegetation and overall appearance sufficient to justify the *brownfield* classification. The third image displays a greenhouse, which ChatGPT incorrectly classified as a *construction* site. These examples highlight the nuances in classification that can lead to discrepancies between human and LLM annotations. Finally, the third group of images illustrates the types of samples that would have been passed to the expert annotator according to strategy S_3 —meaning the LLM and the nonexpert annotator had previously disagreed. In all three cases, ChatGPT would have produced a false positive for the *brownfield* class. In the first case, the model correctly identifies the empty parking lot but misinterprets the vegetation as “overgrown.” In the second case, ChatGPT fails to notice the visible tire tracks in the sand that indicate recent human activity and a potential construction site. Although the model mentioned the vegetation in the bottom right corner, it seems the expert considered the lack of vegetation on the left side combined with the tire tracks as evidence for a construction site. The last example displays a large warehouse. ChatGPT argues that the roof is evidence of a brownfield, but both human annotators disagreed, possibly due to trucks parked on the top side. Overall, it can be concluded that the reasoning provided by the LLM is often logical, but its focus and prioritization of different visual features can sometimes be incorrect. It remains to be investigated to what degree such disagreements can be mitigated by providing more precise prompts. Fine-tuning the LLM given the human expert’s feedback is out of scope and considered for future research.

VII. DISCUSSION

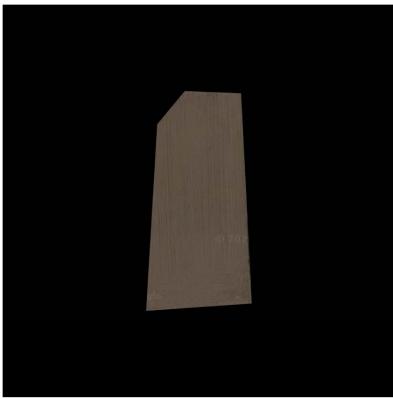
Infallible expert assumption: Some remarks need to be made to put the results into context. First, it is valid to question the gold-standard of this study; the expert annotation. We assume that the expert always assigns the correct label and, thus, define expert agreement as desirable annotator behavior. However, this assumption most likely does not always hold true. Domain experts are unlikely to be error- and bias-free, especially in this study where they had less tools available for their annotations than usual. Whenever possible, it is advisable to consult multiple expert annotators to create gold-standard annotations.

Single-label assumption: Another assumption to be challenged is the single-label convention. We assumed that only one usage class can be attributed to one parcel. In rare cases, this assumption proved to be wrong, meaning a multilabel convention should be reconsidered for future work. Fig. 8 shows two examples of satellite images containing two classes.

A likely reason for this misconception might be that the annotation convention is defined under the assumption that only one parcel is visible on each image. In reality, however, small inconsistencies in the pipeline or the underlying parcel data could lead to multiple parcels with different usage being shown in the same image.⁵

⁵ Parts of a neighboring parcel with a different usage type could be visible on some images due to buffers defined when extracting the images.

Example images with a ChatGPT certainty of 1 or 2 (1.1 – 1.3)

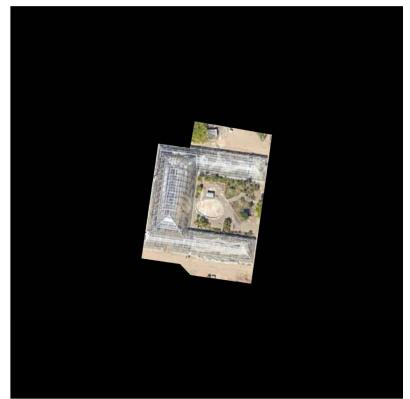


GPT	Expert	Non-exp.
BR	BR	BR

GPT	Expert	Non-exp.
BR	BR	CO

GPT	Expert	Non-exp.
BR	CO	CO

Example images with a ChatGPT certainty of 4 but on which expert and LLM disagreed (2.1 – 2.3)

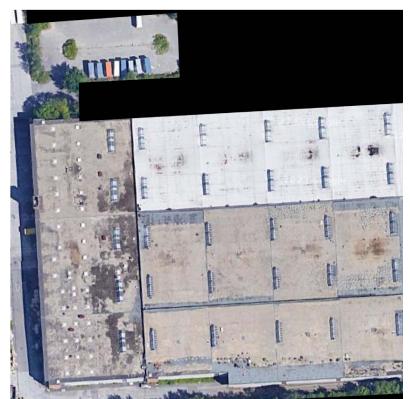


GPT	Expert	Non-exp.
CO	AC	BR

GPT	Expert	Non-exp.
CO	BR	CO

GPT	Expert	Non-exp.
CO	AC	AC

Images with ChatGPT certainty below 4 on which LLM and non-expert human annotator (RA1) disagreed (3.1 – 3.3)



GPT	Expert	Non-exp.
BR	AC	AC

GPT	Expert	Non-exp.
BR	CO	CO

GPT	Expert	Non-exp.
BR	AC	AC

Fig. 7. Inspection of disagreement cases between certain annotator types. Sources: Google [24] for images 1.1, 1.2, 2.2, 3.3; GeoBasisDaten/BKG (2024) [9] for 1.3, 2.3, 3.1 and Bing [39] for 2.1, 3.2. Certainty scores are from the “GPT long 0.2” configuration.

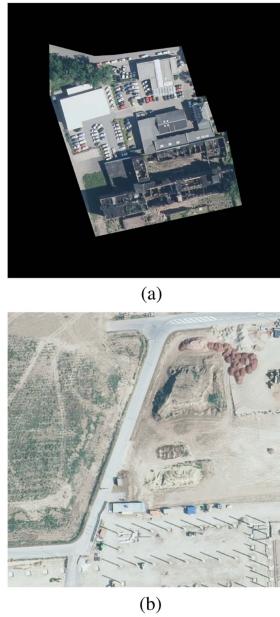


Fig. 8. Multiclass satellite images. Source: GeoBasisDaten/BKG (2024). (a) Brownfield and active. (b) Brownfield and construction.

Task-specific expertise: Considering the outliers in both the ML engineer and the layperson groups from Section V-A, we need to reevaluate what type of expertise is crucial for success in the annotation task. It appears that assumed domain expertise might not be the key determinant of annotation data quality. Instead, task-specific aptitude of each individual annotator plays a more important role. Further investigation revealed that ML Engineer 2 (ML2) was not as closely involved in previous projects, resulting in a lack of task-specific knowledge compared to ML Engineer 1 (ML1) and 3 (ML3). Conversely, Layperson 3 (LP3) had extensive prior experience with satellite image classification, which was not disclosed during data collection. Despite this important finding, we chose not to regroup annotators based on their *true* task-specific expertise. We believe that thorough reporting as presented offers greater scientific value. Annotator quality cannot be known *a priori* and is not necessarily inferred from simple classifications like “layperson.” This approach emphasizes the importance of assessing individual capabilities rather than relying solely on assumed expertise. A “good” annotator can be an expert, a well instructed layperson or an LLM. While it could be argued that this study should have included a larger number of research assistants and laypersons to obtain a more robust majority vote, we contend that neither type of human annotator is likely to be available in large numbers for the most annotation efforts. Future work could consider incorporating crowdworkers from platforms such as Amazon MTurk as another potential annotator type.

Limitations of using LLMs for annotation: Regarding ChatGPT annotations, we surprisingly observe a higher overall agreement with the expert when providing less task-specific context (in a shorter prompt) but a better *brownfield* detection rate for the long prompt. This difference in behavior by the LLM is worth to be discussed. Ultimately, the behavior could have been caused by different prioritization instructions in the two prompts. The long prompt (see Fig. 12 in the Appendix) specifically instructed the model to consider *construction* as first and *brownfield* as second priority. In contrast, the short

prompt instructed the model to report the most likely class. However, the fact that the long prompt performed worse for its first priority (i.e., *constructions*) compared to the short prompt indicates that multiple effects might be at play. Our results highlight the general need for further research around prompt engineering. Prompts for LLM annotation need to be carefully engineered, depending on the most important classes of interest and error types to avoid, like the misclassification of *brownfields* in this study. Here, experimenting with few-shot prompting could yield valuable insights and potentially improved annotation quality. Other minor modifications of the prompt such as omitting or modifying the reasoning and confidence score could also affect the LLMs responses. In addition, the models provided by OpenAI should be compared to LLMs from other providers. Aside from prompt engineering, anyone intending to use LLMs for image annotations should be aware of several challenges. First, LLMs are known to hallucinate and ChatGPT is not immune to it. The reported reasonings in Table VII in the Appendix show no major issues. However, around 1% of the responses by the API are provided in an incorrect format or hallucinated and are, therefore, coded as NAs. Second, API restrictions might apply. While the API calls themselves are fast, a prompt token limit required us to send images in small batches. Third, using automated annotation tools might lead to a reinforcement of existing biases or an emergence of new ones. Humans may tend to over-rely on automatically generated annotations. Automated annotators, just like human annotators, could be subject to well-known mitigators of data quality such as ordering effects or confirmation bias. Lastly, licensing of LLM output and usage is an important consideration. For open-source LLMs, commercial usage might be restricted. For ChatGPT, Open AI’s terms of service apply.

VIII. CONCLUSION

In this study, we conduct an extensive, in-depth comparison of annotation quality across various annotator types for an expert domain image annotation task. We extend the existing literature in remote sensing by comparing 7 different types of human and automated annotators. With that, our study contributes to a discussion about who can be considered an expert for an annotation task and how the expertise of interest can be identified and characterized. In addition, we evaluate combining automated and human annotators in a hybrid annotation workflow. We show that ChatGPT’s self-reported certainty score can be used as a proxy for image-level difficulty and propose a routine to select difficult samples for reannotation by human experts based on this score. Compared to annotating all images with human experts, a hybrid annotation strategy combining expert and LLM annotations would have resulted in a cost reduction of around 50%. Label quality, measured as agreement with the expert, remains high with around 95%. We also highlight the need for calibration of this approach. An ablation study demonstrates that combining automated and human annotators is also feasible without LLMs. As implications from our results, we point out several ways for future research into leveraging LLMs for data annotation and discuss its limitations. Extensive considerations should go into the design of task instructions, guidelines, and prompts. Building on our results, we recommend testing multiple human and automated annotators on a smaller

set of data and carefully evaluating the results in order to understand the differences and anticipated quality loss. Ultimately, one could argue that human annotators now need to demonstrate the value they add over LLMs. However, this work provides evidence that strategically combining human and LLM annotations can lead to efficiency gains, presenting a promising path forward.

APPENDIX



(a)



(b)



(c)

Fig. 9. Illustration of the images from the three sources for the same parcel. Due to variety in provider's actuality of the images, classes are not guaranteed to be consistent across the three sources Google, Bing, and [9]. (a) Bing (active). (b) GeoBasisDaten/BKG (2024) (active). (c) Google (construction).

Additional information on regression parameters: For the three human annotator models, we add annotation time to the independent variables. Since annotation time is not correlated at all within one annotator type ($| \text{correlation coefficients} | < 0.03$) we use the median time. For research assistants and laypersons, we add order as explanatory variable. Order is always correlated strongly by > 0.96 and we compute the average order. The ChatGPT responses provided a certainty score on the same scale as the experts, which is additionally used in the ChatGPT regression models. Similarly, the CL annotations provided a label quality score ranging from 0 to 1, which is added to the respective model.

Algorithm for strategy S_2 in Section VI

Algorithm 2: Hybrid Strategy S_2 .

```

1: Start with a pool of unlabeled samples  $X$ .
2: for each sample  $x$  in  $X$  do
3:   compute  $L_{GPT}(x)$  and  $\theta(x)$ 
4:   if certainty score  $\theta(x) == 4$  then
5:      $L_{S_2}(x) = L_{GPT}(x)$ 
6:   else
7:     compute expert label  $L_{expert}(x)$ 
8:     set  $L_{S_2}(x) = L_{expert}(x)$ 
9:   end if
10: end for
```

Algorithms for ablation study in Section VI

Algorithm 3: Ablation for Hybrid Strategy S_2 With CL Instead of LLM.

```

1: Start with a pool of unlabeled samples  $X$ .
2: for each sample  $x$  in  $X$  do
3:   compute  $L_{CL}(x)$  and  $\theta(x)$ 
4:   if certainty score  $\theta(x) >= 0.99$  then
5:      $L_{S_2}(x) = L_{CL}(x)$ 
6:   else
7:     compute expert label  $L_{expert}(x)$ 
8:     set  $L_{S_2}(x) = L_{expert}(x)$ 
9:   end if
10: end for
```

Algorithm 4: Ablation for Hybrid Strategy S_3 With CL Instead of LLM.

```

1: Start with a pool of unlabeled samples  $X$ .
2: for each sample  $x$  in  $X$  do
3:   compute  $L_{CL}(x)$  and  $\theta(x)$ 
4:   if certainty score  $\theta(x) >= 0.99$  then
5:      $L_{S_3}(x) = L_{CL}(x)$ 
6:   else
7:     compute  $L_{non_{expert}}(x)$ 
8:     if  $L_{CL}(x) == L_{non_{expert}}(x)$  then
9:        $L_{S_3}(x) = L_{CL}(x)$ 
10:    else
11:      compute expert label  $L_{expert}(x)$ 
12:      set  $L_{S_3}(x) = L_{expert}(x)$ 
13:    end if
14:  end if
15: end for
```

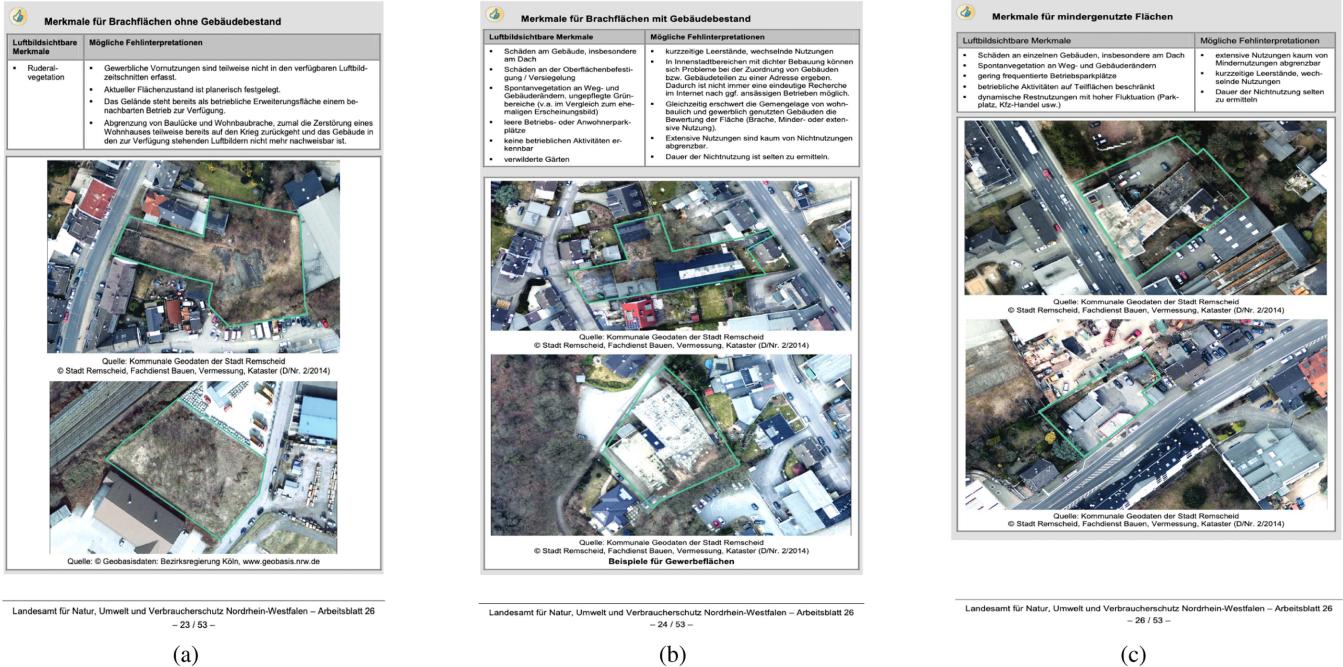


Fig. 10. Official instructions for the identification of *brownfields* [31]. Annotators are additionally instructed to consider *construction* as first and *brownfield* as second priority (same as in the long prompt in Fig. 12 in the Appendix). (a) Undeveloped. (b) Developed. (c) Underutilized.

Context:
You are an expert in annotating satellite images for a machine learning problem and you will be asked to select the most appropriate label to an image for an important research task.

Definitions/rules:
The only 3 options to select of are Brownfield, Actively used Land and Construction.

Priorities:
If you are uncertain, select the most likely one.

Task:
In addition provide a reasoning for your selection and a certainty score from 1 (very uncertain) to 4 (very certain). Separate the label, reasoning and certainty by comma in your response. separate the label and the reasoning by comma in your response.

Model parameters for training CNN-Classifier.

TABLE VI
PARAMETER SPECIFICATIONS FOR MODELS REPORTED IN TABLE III

Parameter	Description
Backbone	resnet50
Validation	CV with 5 folds.
Training size	2370
Test fold size	585
Learning rate (lr)	0.0001
Epochs	20
Batch size	16
Pre-trained	We instantiate models with weights from TorchGeo [55] licensed under an MIT license (pretrained on remote sensing images)
Decay	0.0001
Data Augmentation	
CenterCrop	128
Normalize	True
RandomGrayscale	0.2
RandomHorizontalFlip	0.2
RandomRotation	(0, 180)
Resize	(256, 256)

Fig. 11. GPT short prompt.

Context:

You are an expert in annotating satellite images for a machine learning problem and you will be asked to select the most appropriate label to an image for an important research task.

Definitions/rules:

I am now going to provide you with a short definition of the three classes and how to distinguish between them. The first class is Brownfield. We define Brownfields as properties that have been abandoned for a longer period of time after commercial or industrial use, and therefore have the potential for new uses. Brownfields can be divided into 3 subclasses: undeveloped, developed or underutilized. Undeveloped brownfield sites refer to areas with idle land. Typically, these are covered with inhomogeneous vegetation or soil. Parts of the surface could also be sealed by concrete. Developed brownfield sites refer to areas with abandoned or broken down properties. These sites are often characterized by vacant and not well-maintained buildings, damaged roofs and pavement – without any visible sign for meaningful business activities. Underutilized brownfields often display cars, pieces of scrap or other objects scattered chaotically without clear signs for business activity. The second class is Construction and it is important not to confuse Brownfields and Construction. The important distinction between a Brownfield and a construction site is that the former displays human activity, such as earth moving activities, installation and deinstallation, visibility of machinery or construction materials or partly constructed buildings. The third class is actively used land and includes all properties that are not under construction and that show clear human commercial or industrial usage, independently of any specific sector or type of business operation.

Priorities:

First priority: If there is evidence for construction activity going on, please label such an image as construction. Second priority: Some images might show a mix of Brownfield and active characteristics in various degrees. Please label the image as brownfield, if a significant part of the property is clearly not actively used.

Task:

You are only allowed to provide a single label for each image: Choose either “Brownfield”, “Actively used Land” or “Construction”. If you are uncertain, select the most likely one. In addition, provide a reasoning for your selection and a certainty score from 1 (very uncertain) to 4 (very certain). Separate the label, reasoning and certainty by comma in your response.

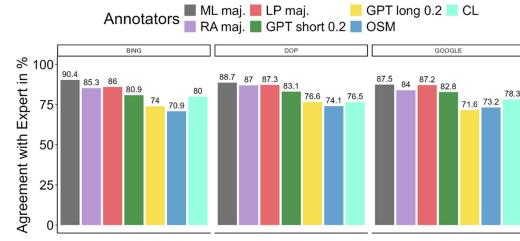


Fig. 13. Agreement with expert by image source.

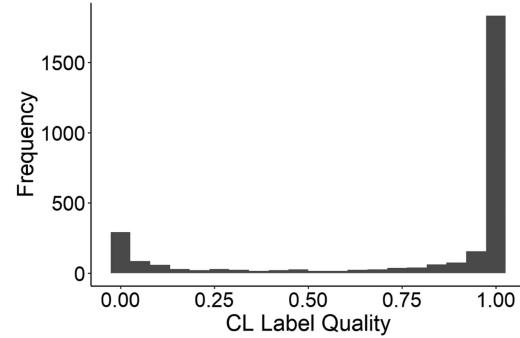


Fig. 14. U-shaped distribution of CL label quality score.

GPT Reasoning for nine example images.

TABLE VII
GPT REASONING FOR SELECTED CASES

Image	GPT Reasoning
1.1	The image shows a large area of idle land with no clear signs of construction or active use
1.2	The image shows an area with no visible signs of active use or construction. The land appears barren with no clear structures, machinery, or materials that would indicate ongoing construction or active commercial/industrial use. The image quality and angle make it difficult to discern specific details, but the lack of activity suggests it is not being actively used.
1.3	The image shows a long stretch of land with a uniform appearance and no visible signs of active use or construction activity, suggesting it could be an undeveloped brownfield site.
2.1	The image shows clear signs of construction activity such as earth moving, scattered construction materials, and what appears to be partially constructed structures or installations.
2.2	The image shows evidence of earth moving activities with visible tracks of heavy machinery and piles of materials, which are indicative of ongoing construction work.
2.3	The image shows a structure with exposed steel reinforcements which is indicative of ongoing construction activity, as well as what appears to be construction materials or debris scattered around the site.
3.1	The image shows a property with no visible signs of active use or construction. The area appears to be a parking lot with overgrown vegetation and no cars, suggesting it is not in active use.
3.2	The image shows a site with no visible active construction or commercial/industrial activity, and there are signs of idle land with inhomogeneous vegetation or soil.
3.3	The image shows a large building with a roof that appears to be in disrepair, with no visible signs of ongoing construction or active use.

Fig. 12. GPT long prompt.

REFERENCES

- [1] T. D. Aguda, S. Siddagangappa, E. Kochkina, S. Kaur, D. Wang, and C. Smiley, "Large language models as financial data annotators: A study on effectiveness and efficiency," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval.*, 2024, pp. 10124–10145. [Online]. Available: <https://aclanthology.org/2024.lrec-main.885>
- [2] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *Proc. Int. joint Conf. neural Netw.*, 2020, pp. 1–8.
- [3] L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," *AI Mag.*, vol. 36, no. 1, pp. 15–24, Mar. 2015. [Online]. Available: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564>
- [4] J. Beck, "Quality aspects of annotated data," *ASzA Wirtschafts- und Sozialstatistisches Archiv*, vol. 17, no. 3, pp. 331–353, Dec. 2023, doi: [10.1007/s11943-023-00332-y](https://doi.org/10.1007/s11943-023-00332-y).
- [5] J. Beck, S. Eckman, R. Chew, and F. Kreuter, "Improving labeling through social science insights: Results and research agenda," in *Proc. HCI Int. Late Breaking Papers: Interacting eXtended Reality Artif. Intell.*, 2022, pp. 245–261.
- [6] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. V. D. Oord, "Are we done with imagenet?," 2020, *arXiv:2006.07159*.
- [7] Bundesamt für Kartographie und Geodäsie (BKG), "Digitales Basis-Landschaftsmodell Ebenen (Basis-DLM-Ebenen)," Accessed: Sep. 16, 2024. [Online]. Available: <https://gdz.bkg.bund.de/index.php/default/digitales-basis-landschaftsmodell-ebenen-basis-dlm-ebenen.html>
- [8] Bundesamt für Kartographie und Geodäsie (BKG), "Flurstücksinformationen Deutschland (FS-DE)," Accessed: Sep. 16, 2024. [Online]. Available: <https://gdz.bkg.bund.de/index.php/default/flurstuecksinformationen-deutschland-fs-de.html>
- [9] Bundesamt für Kartographie und Geodäsie (BKG), "Digitale Orthophotos Bodenauflösung 20cm (DOP20)," Accessed: >Sep. 16, 2024. [Online]. Available: <https://gdz.bkg.bund.de/index.php/default/digitale-geodaten-digitale-orthophotos-digitale-orthophotos-bodenauflösung-20-cm-dop20.html>
- [10] Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit (BMUB), "Klimaschutzplan 2050," 2016. [Online]. Available: https://www.bmwi.de/Redaktion/DE/Publikationen/Industrie/klimaschutzplan-2050.pdf?__blob=publicationFile&v=1
- [11] J. Chen and J. Mueller, "Quantifying uncertainty in answers from any language model and enhancing their trustworthiness," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 5186–5200.
- [12] L. Chen, A. Perez-Lebel, F. Suchanek, and G. Varoquaux, "Reconsidering LLM uncertainty from the grouping loss perspective," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, pp. 1567–1581, 2024.
- [13] Y. Chen and J. Joo, "Understanding and mitigating annotation bias in facial expression recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14960–14971. [Online]. Available: <https://ieeexplore.ieee.org/document/9710276/>
- [14] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
- [15] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [16] M. Desmond, E. Duesterwald, K. Brimjoi, M. Brachman, and Q. Pan, "Semi-automated data labeling," in *Proc. NeurIPS 2020 Competition Demonstration Track*, 2021, pp. 156–169.
- [17] B. Ding et al., "Is GPT-3 a good data annotator?," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 11173–11195.
- [18] R. Dominguez-Olmedo, M. Hardt, and C. Mendler-Dünner, "Questioning the survey responses of large language models," 2024, *arXiv:2306.07951*.
- [19] A. Dumitrache, L. Aroyo, and C. Welty, "Achieving expert-level annotation quality with crowdtruth," in *Proc. BDM21 Workshop*, 2015.
- [20] K. Dürbeck and S. Lippl-Seifert, "Automatische Extraktion von Brachflächen aus Luftbildern mittels eines neuronalen Netzes," *Dresdener Flächennutzungssymposium*, vol. 80, pp. 305–314, 2022.
- [21] S. Gilhuber, R. Hvdingby, M. L. A. Fok, and T. Seidl, "How to overcome confirmation bias in semi-supervised image classification by active learning," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2023, pp. 330–347.
- [22] J.F. Girres and G. Touya, "Quality assessment of the French Open-StreetMap dataset," *Trans. GIS*, vol. 14, no. 4, pp. 435–459, 2010.
- [23] A. Goel et al., "LLMs accelerate annotation for medical information extraction," in *Proc. Mach. Learn. Health*, 2023, pp. 82–100.
- [24] Google, "Google Maps," 2024. Accessed: Sep. 16, 2024. [Online]. Available: <https://www.google.com/maps>
- [25] M. Haklay, "How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets," *Environ. Plan. B: Plan. Des.*, vol. 37, no. 4, pp. 682–703, 2010.
- [26] J. He, J. V. Ossenbruggen, and A. P. d. Vries, "Do you need experts in the crowd? A case study in image annotation for marine biology," in *Proc. 10th Conf. Open Res. Areas Inf. Retrieval*, 2013, pp. 57–60.
- [27] H. Heerkens, C. Norde, and B. V. D. Heijden, "Importance assessment of decision attributes: A qualitative study comparing experts and laypersons," *Manage. Decis.*, vol. 49, no. 5, pp. 748–761, May 2011, doi: [10.1108/00251741111130832/full/html](https://doi.org/10.1108/00251741111130832).
- [28] E. Heim et al., "Large-scale medical image annotation with crowd-powered algorithms," *J. Med. Imag.*, vol. 5, no. 03, Sep. 2018, Art. no. 1. [Online]. Available: <https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-5/issue-03/034002/Large-scale-medical-image-annotation-with-crowd-powered-algorithms/10.1117/1.JMI.5.3.034002.full>
- [29] C. Kern, S. Eckman, J. Beck, R. Chew, B. Ma, and F. Kreuter, "Annotation sensitivity: Training data collection methods affect model performance," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2023, pp. 14874–14886. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.992>
- [30] J. A. Krosnick, S. Narayan, and W. R. Smith, "Satisficing in surveys: Initial evidence," *New Directions Eval.*, vol. 1996, no. 70, pp. 29–44, Mar. 1996, doi: [10.1002/ev.1033](https://doi.org/10.1002/ev.1033).
- [31] Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen (LANUV), "Leitfaden zur Erfassung von Brachflächen in Nordrhein-Westfalen," 2015. [Online]. Available: https://www.lanuv.nrw.de/fileadmin/lanuvpubl/4_arbeitsblaetter/40026.pdf
- [32] J. Li, "A comparative study on annotation quality of crowdsourcing and LLM via label aggregation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 6525–6529.
- [33] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 2, pp. 32–66, Jun. 2024.
- [34] Y. Li, Z. Zhu, J.G. Yu, and Y. Zhang, "Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590–10603, Dec. 2021.
- [35] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 6, 2018, Art. no. e1264.
- [36] Z. Li, L. Lin, C. Zhang, H. Ma, and W. Zhao, "Automatic image annotation based on co-training," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
- [37] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *Int. J. Remote Sens.*, vol. 28, no. 5, pp. 823–870, 2007.
- [38] K. Maaß, U. Trautwein, C. Gresch, O. Lüdtke, and R. Watermann, "Intercoder-reliabilität bei der berufscodierung nach der ISCO-88 und validität des soziökonomischen status," *Zeitschrift für Erziehungswissenschaft*, vol. 12, no. 2, pp. 281–301, Jun. 2009. [Online]. Available: <https://link.springer.com/10.1007/s11618-009-0068-0>
- [39] Microsoft Corporation, "Bing Maps," 2024. Accessed: Sep. 16, 2024. [Online]. Available: <https://www.bing.com/maps>
- [40] E. Musi, D. Ghosh, and S. Muresan, "Towards feasible guidelines for the annotation of argument schemes," in *Proc. 3rd Workshop Argument Mining*, 2016, pp. 82–93. [Online]. Available: <https://api.semanticscholar.org/CorpusID:9679538>
- [41] C. Nédellec, P. Bessières, R. Bossy, A. Kotoujansky, and A. P. Marine, "Annotation guidelines for machine learning-based named entity recognition in microbiology," in *Proc. ACL Workshop Data Text Mining Integrative Biol.*, 2006, pp. 40–54.
- [42] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, 2021.
- [43] C.G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 1)*, 2021, pp. 1–13.
- [44] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation," in *Proc. Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 557–566.
- [45] OpenAI, "OpenAI Developer Platform," Accessed: Sep. 16, 2024. [Online]. Available: <https://platform.openai.com/docs/guides/vision>

- [46] OpenStreetMap contributors, "Planet OSM," Accessed: Sep. 16, 2024. [Online]. Available: <https://planet.osm.org>
- [47] L.P. Osco, E. L. d. Lemos, W. N. Gonçalves, A. P. M. Ramos, and J. Marcato Jr., "The potential of visual ChatGPT for remote sensing," *Remote Sens.*, vol. 15, no. 13, 2023, Art. no. 3232.
- [48] Overpass Turbo, "Overpass turbo: A web-based data querying tool for Openstreetmap," Accessed: Sep. 16, 2024. [Online]. Available: <https://overpass-turbo.eu/>
- [49] N. Pangakis, S. Wolken, and N. Fasching, "Automated annotation with generative AI requires validation," 2023, *arXiv:2306.00176*.
- [50] J. Roberts, T. Lüdecke, S. Das, K. Han, and S. Albanie, "GPT4GEO: How a language model sees the world's geography," 2023, *arXiv:2306.00020*.
- [51] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, "Position: Mission critical-satellite data is a distinct modality in machine learning," in *Proc. 41st Int. Conf. Mach. Learn.*, pp. 42691–42706, 2024.
- [52] K. Saito, D. Kim, and K. Saenko, "OpenMatch: Open-set semi-supervised learning with open-set consistency regularization," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 25956–25967.
- [53] A.K. Singh, S. Devkota, B. Lamichhane, U. Dhakal, and C. Dhakal, "The confidence-competence gap in large language models: A cognitive study," 2023, *arXiv:2309.16145*.
- [54] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Annu. Conf. neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.
- [55] A.J. Stewart, C. Robinson, I.A. Corley, A. Ortiz, J. M. L. Ferres, and A. Banerjee, "Torchgeo: Deep learning with geospatial data," in *Proc. 30th Int. Conf. Adv. Geographic Inf. Syst.*, 2022, pp. 1–12.
- [56] Cleanlab, "The standard data-centric AI package for data quality and machine learning with messy, real-world data and labels," Accessed: Sep. 16, 2024. [Online]. Available: <https://github.com/cleanlab/cleanlab>
- [57] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label studio: Data labeling software (2020–2024)," 2024. Accessed: Sep. 16, 2024. [Online]. Available: <https://github.com/HumanSignal/label-studio%7D>
- [58] P. Törnberg, "ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning," 2023, *arXiv:2304.06588*.
- [59] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "Open-StreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 184–199, Mar. 2021.
- [60] P. Wang and N. Vasconcelos, "Towards professional level crowd annotation of expert domain data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 3166–3175. [Online]. Available: <https://ieeexplore.ieee.org/document/10204827/>
- [61] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural network and its application in image classification: A survey," *Opt. Eng.*, vol. 58, no. 4, 2019, Art. no. 040901.
- [62] X. Wang et al., "Accurate label refinement from multiannotator of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 4700413.
- [63] Y. Yang, O. Agarwal, C. Tar, B.C. Wallace, and A. Nenkova, "Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 1471–1480.
- [64] F. Ye et al., "Benchmarking LLMs via uncertainty quantification," 2024, *arXiv:2401.12794*.
- [65] D. Yu, L. Li, H. Su, and M. Fuoli, "Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis," *International Journal of Corpus Linguistics*, vol. 29, vol. 4, pp. 534–561, 2024.
- [66] Z. Yuan et al., "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3078451.
- [67] S. Yun, S. J. Oh, B. Heo, D. Han, J. Choe, and S. Chun, "Re-labeling imangenet: From single to multi-labels, from global to localized labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2340–2350.
- [68] B. Zhang et al., "FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18408–18419.
- [69] Y. Zhang, C. Wei, Z. He, and W. Yu, "GeoGPT: An assistant for understanding and processing geospatial tasks," *Int. J. Appl. Earth Observation Geoinformation*, vol. 131, 2024, Art. no. 103976.
- [70] B. Zhao, Y. Zhong, and L. Zhang, "A spectral-structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 73–85, 2016.
- [71] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2021, Art. no. 5603814.



Jacob Beck received the bachelor's degree from the University of Mannheim, Mannheim, Germany, in 2018, and the master's degree in statistics from the University of Seoul, Seoul, South Korea, in 2021, both in sociology. Since 2021, he has been working toward the Ph.D. degree with the Chair of Statistics and Data Science, LMU Munich, Munich, Germany.

In his research, he focuses on understanding the various drivers of annotated data quality. Generally, he is interested in determining how social science learning can benefit ML applications.



Lukas Malte Kemeter received the bachelor's degree from the University of Mannheim, Mannheim, Germany, in 2018, and the master's degree from the University of Copenhagen, Copenhagen, Denmark, in 2022, both in economics.

Since 2022, he has been a Research Associate with the Fraunhofer Institute for Integrated Circuits, Munich, Germany. His research focuses on data and label efficient learning for machine learning applications.



Konrad Dürrbeck received the M.Sc. degree in management with focus in computer science from the Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, in 2017.

Since 2018, he has been a Research Associate with Fraunhofer Institute for Integrated Circuits, Nuremberg, Germany, and is responsible for the research field of location analysis.



Mohamed Hesham Ibrahim Abdalla is currently working toward the master's degree in computer science with the Technical University of Munich (TUM), Munich, Germany.

He has been a student Assistant with the Fraunhofer Institute for Integrated Circuits, Munich, Germany, since 2022.



Frauke Kreuter received the diploma degree in sociology from the University of Mannheim, Germany, in 1996 and the Ph.D. degree in survey statistics and methodology from the University of Konstanz, Germany, in 2001.

She holds the Chair of Statistics and Data Science with the LMU Munich, Munich, Germany. At the University of Maryland, College Park, MD, USA, she is the Co-Director of the Social Data Science Center (SoDa) and faculty member in the Joint Program in Survey Methodology (JPSM).