Reassessing Active Learning Adoption in Contemporary NLP: A Community Survey

Julia Romberg*♣1, Christopher Schröder*2, Julius Gonsior[3],
Katrin Tomanek[4], and Fredrik Olsson[5]

[1]GESIS – Leibniz Institute for the Social Sciences
[2]Institute for Applied Informatics at Leipzig University (InfAI)
[2]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)
Dresden/Leipzig, Leipzig University
[3]TUD Dresden University of Technology
[4]Independent Researcher
[5]All Ears

Abstract

Supervised learning relies on data annotation which usually is time-consuming and therefore expensive. A longstanding strategy to reduce annotation costs is *active learning*, an iterative process, in which a human annotates only data instances deemed informative by a model. Research in active learning has made considerable progress, especially with the rise of large language models (LLMs). However, we still know little about how these remarkable advances have translated into real-world applications, or contributed to removing key barriers to active learning adoption. To fill in this gap, we conduct *an online survey in the NLP community* to collect previously intangible insights on current implementation practices, common obstacles in application, and future prospects in active learning. We also reassess the perceived relevance of data annotation and active learning as fundamental assumptions. Our findings show that data annotation is expected to remain important and active learning to stay relevant while benefiting from LLMs. Consistent with a community survey from over 15 years ago, three key challenges yet persist—setup complexity, uncertain cost reduction, and tooling—for which we propose alleviation strategies. We publish an anonymized version of the dataset.[1]

1 Introduction

Supervised learning is one of the most common concepts in natural language processing (NLP). By definition, the approach depends on annotated data, which is usually time-consuming to create and therefore expensive, making supervised learning a resource-intensive process.

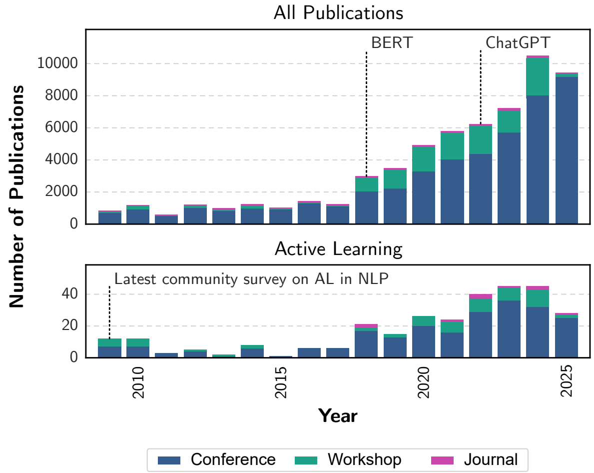


Figure 1: Development of AL publication numbers in *CL venues over the past 15 years (total publication numbers for reference). The methodological details are provided in Appendix A. Note that the numbers for 2025 must be interpreted with caution, as not all data were available in the ACL Anthology at the time of analysis.

A longstanding strategy for minimizing the annotation effort while maintaining model performance is *active learning* (AL; Lewis and Gale, 1994), where a human annotator iteratively provides labels for small batches of data instances deemed informative by a model. The selection of these instances is guided by specific query strategies. After each round, the model is retrained on the newly annotated data to improve task performance. This process continues until a predetermined stopping criterion is met, concluding the annotation process.

Since its beginnings, AL has evolved considerably (Xia et al., 2025), not least because of the gains brought about by large language models (LLMs[2]).

*Equal contribution.
♣Corresponding author: Julia.Romberg@gesis.org
[1]Dataset is available from: https://doi.org/10.7802/2990.

[2]We adopt the definition of LLMs from Rogers and Luccioni (2024), which encompasses both encoders and decoders.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2621–2647

We observe notable surges in the number of AL publications in *CL venues in 2018 and 2022 (see Figure 1), likely due to new opportunities enabled by transformer-based language models and generative AI. These works have found considerable improvements in effectiveness and sample efficiency when integrating LLMs with AL in different scenarios (e.g., Ein-Dor et al., 2020; Margatina et al., 2021; Tonneau et al., 2022; Xiao et al., 2023).

While AL is continuously developed in research, its primary goal remains to enable practical applications. When Tomanek and Olsson last surveyed NLP practitioners on the use of AL in 2009, only 20% stated they had implemented AL to support annotation—although many were aware of AL and its experimentally proven promises. Key barriers to adoption were skepticism about the practical effectiveness and the overhead of implementing annotation interfaces. The few that had adopted AL highlighted sampling complexity, annotator waiting times, and interface design as issues critical to the practical realization. Little do we, as a field, know about how the practical use of AL has altered since then, or whether research advancements have helped address its practical challenges. More than 15 years later, the question of practical adoption thus deserves renewed attention.

Drawing from the literature alone cannot provide a complete answer to this question, as practical use of AL—in contrast to experimental evaluations—is often not documented publicly. Therefore, we devise and conduct an extensive community survey following the example of Tomanek and Olsson (2009). While still uncommon in NLP, community surveys have proven effective for obtaining knowledge unavailable in academic literature (e.g., Zhou et al., 2022; Subramonian et al., 2023; Michael et al., 2023; Blaschke et al., 2024). This way, we can not only assess the assumed transformative impact of LLMs, but also take a step back to re-examine whether fundamental assumptions about the need for data-efficient annotation still hold.

Contributions (1) We collect a dataset of insights on AL and data annotation needs from the NLP community, academia and industry, through a comprehensive survey with 52 questions. (2) We evaluate the responses of 144 participants quantitatively and qualitatively, and compare them to the community survey from 2009. (3) We discuss the results and connect them to literature, describing anticipated future developments in applied AL. (4) We identify three key problems to adoption that persist in contemporary AL and outline possible solutions.

Findings Our descriptive analysis indicates that data annotation remains important and a bottleneck for supervised learning, for which AL is mostly regarded as a relevant method to address this issue. Advances in NLP have led to LLMs becoming the prevalent choice of backbone model in practice, but key barriers to AL adoption from over a decade ago still persist, demanding greater emphasis on reducing the complexity of setup, ensuring cost reduction, and improving annotation tools.

2 Survey Methodology

To fill the knowledge gap between LLM-driven achievements in AL literature and practical implementations, our community survey centers on five research questions. The first two assess assumptions fundamental to AL about the continued relevance of data annotation and AL as a method.

RQ1. What is the current state of data annotation needs in the age of LLMs?

RQ2. How does AL compare to other methods for overcoming data annotation challenges?

We then turn to the main focus of interest, the contemporary implementation of AL with respect to the past (Olsson, 2009) and anticipated future.

RQ3. What do contemporary AL implementations look like in practice?

RQ4. What developments in AL have been observed and are anticipated?

RQ5. How do current perspectives on AL compare with those from 15 years ago?

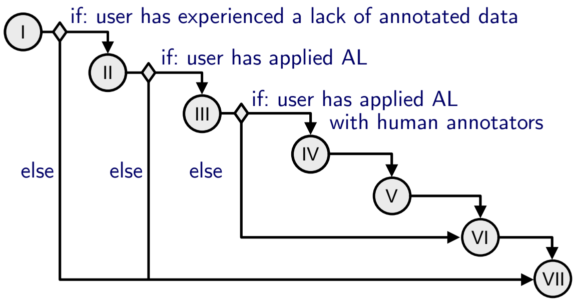
2.1 Descriptive Research Approach

Following previous community surveys in *CL venues (Zhou et al., 2022; Michael et al., 2023; Blaschke et al., 2024), this survey adopts a descriptive research design in order to systematically document experiences within a sample of NLP practitioners. We capture firsthand experiences from academic and industry practitioners relevant to answering the five research questions, including real-world practices, barriers, and further perceptions of AL adoption. Our primary goal with this approach is to understand how and why practitioners use (or avoid) AL in their specific contexts, instead

(a) Question Groups

I.	Relevance of Data Annotation
II.	Overcoming a Lack of Annotated Data
III.	Active Learning: General
IV.	Active Learning: Methodological Setup
V.	Active Learning: Task & Annotation Setup
VI.	Trends in Active Learning for NLP
VII.	Background Information

(b) Flow



(c) Paths

Path Name	Groups	#Questions
Full	I–VII	52
Active learning without human annotators	I–III, VI–VII	32
No active learning	I–II, VII	22
Never experienced a lack of annotated data	I, VII	8

Figure 2: This diagram illustrates (a) the names and sequence of question groups in the survey, (b) the logical flow between these question groups, including branching points and conditions, and (c) the possible paths that respondents take through the survey, ordered by the number of questions in descending order.

of making universal claims about AL's effectiveness. Notably, descriptive research methodology provides valuable documentation, even if it does not generalize to the entire target population. The practitioners' responses provide empirical evidence that has at most been theoretically discussed, but not systematically documented before. Our findings can be interpreted as valuable documentation of how AL is actually employed in practice, with important implications for future research.

2.2 Questions and Branching

Our survey comprises 52 questions over 7 logical groups (Figure 2a), which we will refer to by question group I to VII in the following. The survey includes both predefined and free-form response options. We refer to the latter as *other* responses, and assess them through qualitative coding, involving normalization and grouping of the inputs.

We use branching logic (Figure 2b) to guide participants through the survey. The main criteria here are having (1) experienced a lack of annotated data, (2) applied AL in either setting (i.e., simulation or practical), and (3) used AL in practical settings with human annotators. Questions are automatically skipped if participants are ineligible to answer based on prior responses. For example, participants who have never used AL will not be shown groups III–VI (see Figure 2c for different user groups). Thereby we aim to minimize the required effort for completing the survey, which is more appreciative of the participants' time, and also has been shown to positively affect the completion rates (Liu and Wronski, 2018). The full survey is provided in Appendix C, including information on each question's format (i.e., free-form, multiple or single choice).

2.3 Target Audience and Distribution

The survey is targeted at the NLP domain and the minimum eligibility for participation is basic knowledge of supervised learning for NLP.

We distributed the survey through various channels aiming to reach a broad audience: (1) mailing lists (ACL, ELRA corpora, ELRA SIGUL, Natural Language Processing DC, tada.cool); (2) personalized email to a manually curated list of 601 individuals who co-authored papers on AL at major *CL venues between 2009 and 2024, and 9 additional personal contacts; (3) common social media channels (LinkedIn, Bluesky, Twitter/X, Hugging Face Posts); (4) outreach to annotation tool developers and providers, asking them to share the survey with their users. Supplementary details on distribution and implementation are provided in Appendix B.

The survey was open online to voluntary participants for 6 weeks, from December 15th, 2024, to January 26th, 2025. Calls for participation were shared in three waves: the initial invitation at the start of the survey period, a first reminder after two weeks, and a second reminder after four weeks.

2.4 Consent, Data Usage, and Privacy

We implemented measures to ensure compliance with the EU General Data Protection Regulation. The survey was conducted anonymously, and participants consented to an anonymized release of their input. See Appendix B.4 for details.

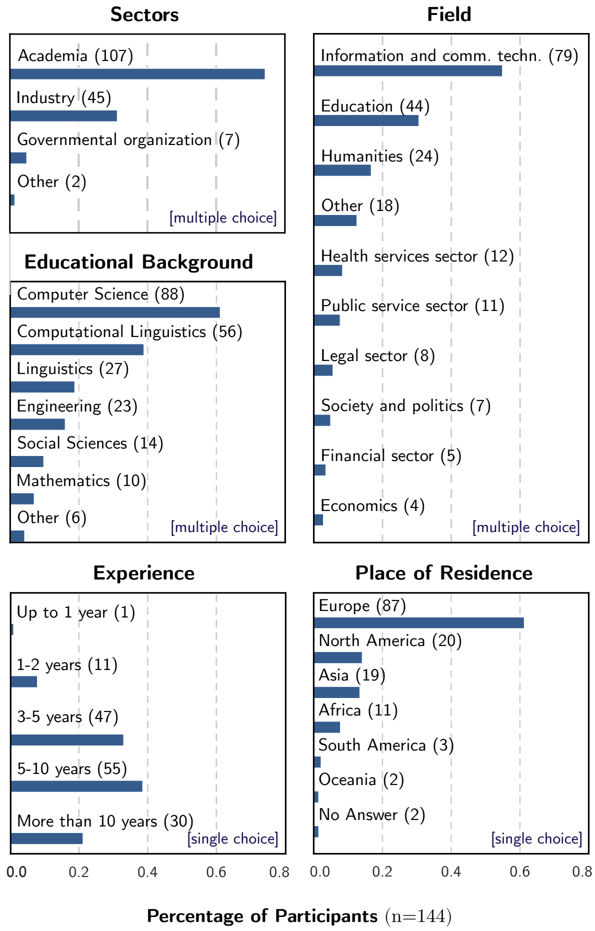


Figure 3: Exploratory histograms of participants' work sectors, the field they are working in, educational background, work experience in machine learning/NLP, and primary place of residence (cf. VII.1-5). The y-axis of each histogram represents the response options for the survey question, while the x-axis shows the proportion of participants who chose each response option.

3 Results

We first present the participant sample and then analyze the responses to RQ1–RQ4, deferring the discussion of RQ5 to Section 5.

Table 3 in Appendix D lists the response distributions for all 52 survey questions. In the remainder, we link specific arguments in the text to corresponding questions in Table 3 by question group and number (e.g., I.1 for the first question).

3.1 Participant Statistics

Among 171 persons that navigated to the first question group, 144 completed the survey.[3] This corresponds to a completion rate of 84% of those that entered the content part of the survey. Incomplete responses were excluded from analysis.

Figure 3 presents a demographic summary of the 144 participants. Most work in academia, but a substantial portion is employed in industry. Information and communications technology clearly leads the field of work, while the educational background is primarily in computer science and computational linguistics. Overall, the participants are very experienced in NLP, with the majority having between 3 to 10 years of experience. For recruiting participants, mailing lists were the most effective referral source (51%, cf. VII.7).

A certain degree of selection bias is evident, and should be kept in mind when interpreting our results. Most notable is a geographical overrepresentation from Europe, despite considerable efforts to reach a diverse audience (see Appendix B; this limitation is also encountered by, e.g., Højer et al. (2025) despite a global recruitment strategy). Our targeted recruitment of authors may have introduced additional bias, but was deemed necessary to reach sufficient response rates on AL questions. Appendix D.1 addresses potential overestimation.

3.2 Data Annotation in Times of LLMs

First, we assess whether the community perceives a persistent need for data annotation (RQ1). To capture perceived shifts in the relevance of annotated data, we ask those 138 respondents (out of 144) who had previously encountered difficulties due to a lack of annotated data (cf. I.1). We are interested in potential changes in difficulties given recent advancements in NLP.

Figure 4 illustrates the attitude on five questions about the relevance of data annotation currently. Overall, participants largely agree that many problems will still require supervised learning (80%). Annotated data remains a limiting factor here; in general (94%), for certain languages (75%), and for problems of a certain complexity (91%). Moreover, for their tasks, only few participants agree that generative AI for training data synthesis offers a viable solution (30%).

Answer to RQ1: Supervised learning remains important in the age of LLMs. Data annotation is still a bottleneck, especially in demanding scenarios.

3.3 Overcoming Data Annotation Challenges

Next, we ask how AL is perceived for overcoming data annotation challenges compared to alternative methods (RQ2). 120 out of the 138 respondents

[3]Compared to other surveys (Zhou et al., 2022; Subramonian et al., 2023; Michael et al., 2023; Blaschke et al., 2024), 144 full responses place us in the mid-range of participation.

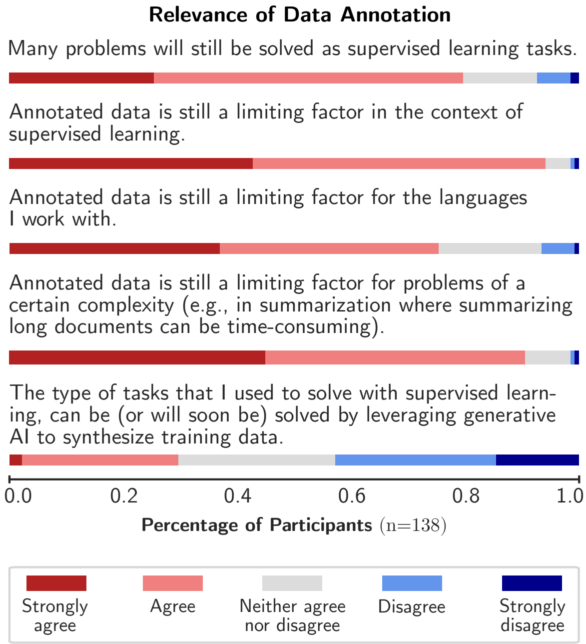


Figure 4: Respondents' assessment of the relevance of data annotation in context of recent advancements in NLP using a 5-point Likert scale (cf. I.3–7).

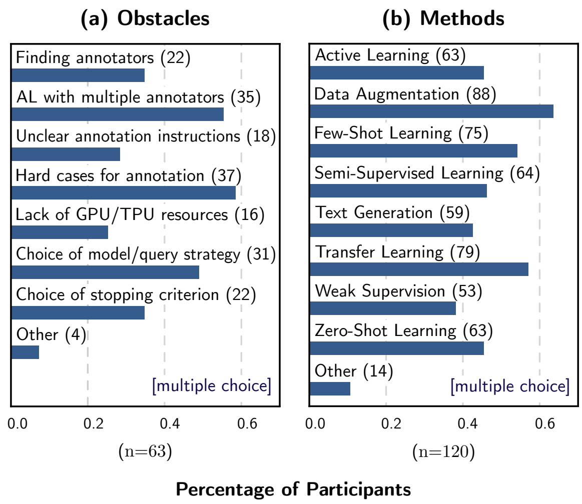


Figure 5: Obstacles especially faced in the application of AL (a) and participants' general experience with computational methods for a lack of annotated data (b). The y-axis represents the response options for the survey question, while the x-axis shows the proportion of participants who chose each response option.

who were eligible for RQ1 had used computational methods to answer this (cf. II.1).

Adoption of active learning AL has been used by over half of the respondents (a total of 63; cf. II.2), and the primary motivation was obtaining annotated data at minimal cost (87%). Further goals included gaining practical experience (46%), identifying difficult examples in annotation guidelines (24%), and improving data quality or using AL for research purposes (8%; cf. III.2).

When asking why participants had not chosen AL, 25% had never heard of it, and 49% cited insufficient expertise. There were accompanying methodological concerns, such as the expected implementation overhead (37%), the lack of (or unawareness about) suitable annotation tools (32%), the difficulty of estimating effectiveness upfront (12%), and sampling bias (18%). 9% found it unsuitable due to specific project requirements and 14% doubted AL's effectiveness at all (cf. II.5).

Despite these concerns, 54% of participants who had not yet used AL consider its application in future projects and further 38% indicated they are merely uncertain due to a lack of knowledge. Only one participant would explicitly not use AL (reasoning that it does not work well enough in practice; cf. II.8), while four made its use dependent on the specific circumstances (16%; cf. II.7).

Alternative methods The current users of AL also report obstacles in data annotation with AL, illustrated in Figure 5a, which may lead them to opt for alternatives instead. The most pronounced challenge is handling hard annotation cases, followed by the involvement of multiple annotators. The third major hurdle is selecting an appropriate model and query strategy. In the other responses, the need for open, robust, and user-friendly AL tools was emphasized (cf. III.3).

The use of computational alternatives to AL is widely dispersed across participants (see Figure 5b), but we observe that only data augmentation, transfer learning, and few-shot learning have been used more frequently than AL in our sample. In general, the alternative methods were considered successful in overcoming a lack of annotated data in slightly over half of the cases (57%, cf. II.4). In comparison, recent AL-based annotation projects were considered successful in 91% and effective in 67% of cases (see Section 3.4 for more details).

Answer to RQ2: The strong openness to and perceived success of AL suggest its relevance for data annotation despite available alternatives.

3.4 Contemporary Active Learning

Moreover, we ask what contemporary setups of applied AL look like (RQ3). This section evaluates

Figure 6: Distribution of the reported combinations of model and query strategy.

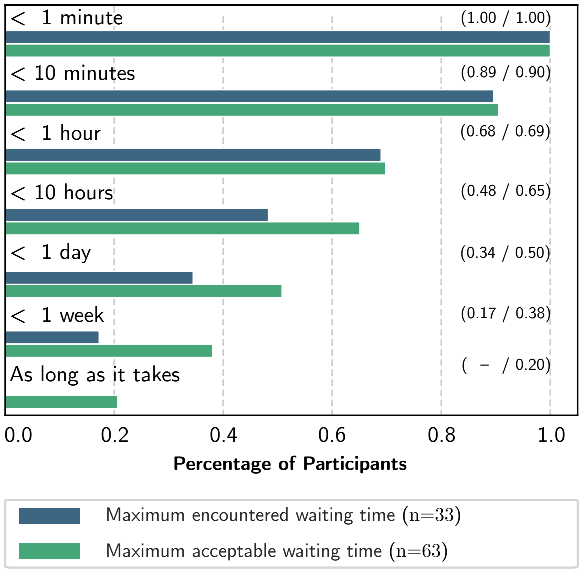


Figure 7: Waiting times between two annotation cycles encountered in practical projects (cf. IV.4) versus the maximum time respondents would accept (cf. III.4).

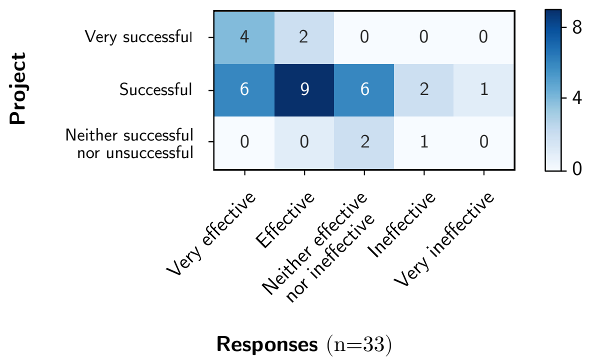


Figure 8: The reported project success broken down by effectiveness of AL. Unselected options for project success are omitted (unsuccessful, very unsuccessful).

33 AL projects[4] with human annotators that were conducted within the last five years (from 2020 onwards; cf. III.5, IV.1).

Models, query strategies, and stopping criteria Figure 6 provides an overview of models and query strategies used. The majority of projects relied on LLMs, with a clear preference for smaller models such as BERT. Notably, only two projects employed larger language models (in the context of our survey with 1B parameters and above). Uncertainty sampling emerged as the dominant query strategy, despite the wide range of alternatives.[5] Among the other options, we observe the tryout of multiple models and query strategies, as well as AutoML libraries for model selection (cf. IV.2,3). For stopping the AL process, most projects halted upon budget depletion (39%), after evaluation of the model on a held-out gold standard (30%) or via human assessment (3%). Stopping at a fixed number of iterations or instances is less common (12%). Algorithmic stopping criteria are still rare, used in only 12% of cases (cf. IV.5).

We also inquired regarding encountered waiting times and the maximum acceptable waiting times between annotation cycles, as illustrated in Figure 7. We observe a broad range reported across projects, from under one minute to up to one week. Contrary to our expectations, participants are willing to wait considerable times up to one week, over 50% even as long as one day, suggesting that significant sampling times are neither unexpected nor problematic for them.

Annotation tools Given the availability of annotation tools with AL support, we asked participants about their usage. Interestingly, just over half of the projects used such a tool (17). These include `Argilla` (4), `LabelSleuth` (3), `Prodigy` (2), `ActiveAnno` (1), `ALAMBIC` (1), `ALANNO` (1), `Label Studio` (1), and self-built solutions (4) (cf. V.6,7).

Satisfaction with project Figure 8 illustrates the perceived success of the annotation projects together with the perceived effectiveness of using AL. Overall, most projects were considered successful with an effective AL contribution. At the

[4] There is a near-equal split between academic (16) and industry/government respondents (20, including 3 overlaps).

[5] Zhang et al. (2022) give an overview of query strategies.

same time, especially for successful projects, it can be observed that the perception of AL effectiveness ranges from very high to very low (cf. V.11, V.12).

Among those that reported a reduced effectiveness, the main reasons were the lack of suitable annotation tools (24%; notably, 4 had used a tool with AL support) and the overhead of setup (21%), which connects to insufficient knowledge in retrospect (17%). Performance mismatches contributed, such as poor overall performance (21%), a lack of effort reduction (9%), inaccurate efficiency estimation (9%), and mismatching project-specific requirements (6%). In addition, respondents noted sampling bias (15%) and dataset-model dependency (15%). The other options stress concerns about setup time and effectiveness (cf. V.13). Considering potential differences in terms of practical utility "in the real world" and for NLP researchers, we find that academic and industry/governmental respondents align closely on reasons for reduced effectiveness. Appendix D.2 provides a detailed breakdown across sectors for the entire Section 3.4.

Overall, 28 out of 33 AL practitioners stated they would use AL again in future projects. Two respondents restricted its use to specific scenarios, while another one remained uncertain. Only two respondents would not use AL again, noting that AL does not work well enough in practice (cf. V.14,15).

Answer to RQ3: The reported projects favor LLMs with uncertainty sampling. While project satisfaction is high in general, setup complexity (overhead, efficiency estimates, knowledge), general project risks (performance, effort reduction, sampling bias, model dependency, requirements), and unsuitable tools challenge AL adoption.

3.5 Anticipated Developments

We now turn to developments in AL that respondents anticipate (RQ4), starting with a reflection on four important advancements for NLP from recent years (see Figure 9 in Appendix D). We find that participants particularly agree on the impact of text embeddings (81%), GPU-accelerated computing (67%), and neural networks (59%) on AL implementations. In contrast, the use of parameter-heavy LLMs was more controversial (36% agreement).

We furthermore asked participants, through an open-ended question, about future AL trends in NLP they expect to shape the field. Table 1 shows the results of our qualitative analysis. The majority of responses anticipated the integration of LLMs at various stages of AL. Further emphasis was put on improved instance selection. Responses also hint at expected advancements in tooling or improvements for low-resource languages and subjective tasks.

Leveraging Language Models
1. (Partially) using LLMs as annotators (8)
2. Increased use of LLMs in AL components (7)
3. Data synthesis during or before AL (4)
4. Incorporate language model-based agents (3)
5. Improved usage of embeddings (2)
6. Using small efficient models (2)
Instance Selection
1. More sophisticated query strategies (4)
2. Effectiveness over random sampling (2)
3. More reliable, well-calibrated uncertainty scores (2)
Tooling
1. Improvements in convenience (4)
2. Easier bootstrapping of an AL setup (2)
Other
1. AL for low-resource languages (2)
2. AL for subjective tasks (1)

Table 1: Participant-identified trends in AL for NLP.

Answer to RQ4: Practitioners anticipate that the future of AL will revolve around increased and pragmatic LLM integration. Simultaneously, they call for fundamental improvements in existing query strategies and more accessible tooling.

4 Related Work

Progress in active learning research AL for NLP has substantially grown since its beginnings. Advancements have been reviewed in several literature surveys over time, focusing either on the overall picture (Olsson, 2009; Zhang et al., 2022), or on certain sub-aspects such as deep neural networks and text classification (Schröder and Niekler, 2020), or entity recognition (Kohl et al., 2024). The most recent academic developments are captured by Xia et al. (2025), who surveyed LLM-based AL.

Active learning in the era of LLMs The integration of LLMs and AL has been associated with a number of benefits. Among these are the outstanding performance of LLMs that can be boosted through AL even in demanding real-world data scenarios (Ein-Dor et al., 2020), resulting in a significant reduction of human annotation effort (Shelmanov et al., 2021; Zhao et al., 2020)[6], the addi-

[6]Few studies directly compare the effort reduction in AL between traditional models and LLMs, but Romberg and Escher (2022) show, e.g., that LLMs can halve the manual effort.

tional source of information that pre-training provides in the cold-start of AL (Yuan et al., 2020), and capabilities such as few-shot learning (Margatina et al., 2023; Bayer and Reuter, 2024) and prompt-based strategies (Li et al., 2024). LLMs have thus become the standard backbone models across tasks such as text classification (Ein-Dor et al., 2020; Yuan et al., 2020; Margatina et al., 2021; Galimzianova and Sanochkin, 2024), sequence tagging (Shelmanov et al., 2019, 2021; Luo et al., 2023), text summarization (Tsvigun et al., 2022a; Li et al., 2024; Xia et al., 2024) and machine translation (Zeng et al., 2019; Zhao et al., 2020; Mendonça et al., 2023; Yüksel et al., 2023). Recent studies even suggest to entirely replace the human annotator with LLMs as cost-effective alternatives (Xiao et al., 2023; Zhang et al., 2023; Kholodna et al., 2024; Liang et al., 2024). The reliability of LLM annotators, however, is still under debate. For example, Baumann et al. (2025) showed that with LLM annotators, incorrect statistical conclusions occurred in over a third of cases.

Existing literature on our research questions As anticipated in our motivation for conducting this community survey, the existing literature offers only limited answers to the research questions we address. Related work hardly questions the continued need for data annotation (RQ1) and the relevance of AL as a data-efficient method (RQ2). Our paper contributes by providing evidence to support this common assumption. Regarding contemporary AL setups in real-world use (RQ3), reports from actual AL applications or realistic simulations are rare (Tonneau et al., 2022). That said, a few studies do address challenges that we identified as persisting in practice, such as model-dataset dependency (Shelmanov et al., 2021; Tsvigun et al., 2022c; Yüksel et al., 2023; Jelenić et al., 2023) and unstable performance and distribution mismatch (Li et al., 2024). The literature provides more extensive coverage of trends (RQ4), complementing the list of potential next steps in applied AL. There is initial research on LLMs as annotators (Zhang et al., 2023; Xiao et al., 2023; Liang et al., 2024; Kholodna et al., 2024), AL for few-shot learning or data synthesis (Margatina et al., 2023; Bayer and Reuter, 2024; Zhang et al., 2023), improving AL on low-resource languages through LLMs (Kholodna et al., 2024), cost-efficient LLMs for AL (Galimzianova and Sanochkin, 2024), LLM-determined curricula for stable AL performance (Li et al., 2024), using AL to alleviate LLM hallucinations (Xia et al., 2024), handling human label variation (Wang and Plank, 2023; Van Der Meer et al., 2024) and annotation errors (Weber and Plank, 2023) with AL.

Prior community surveys Academic literature provides insights most of all from simulated experiments, including in literature surveys. Evaluations of practical applications are rare, and more importantly, can only represent the small portion of practical applications that have been documented in a scientific setting. Simulations, in contrast, provide only limited insights into the practical realities (Margatina and Aletras, 2023). As a consequence, important aspects, such as negative results and practical obstacles, are likely missing. For the field of AL, there has only one community survey been conducted as early as 2009 by Tomanek and Olsson. Given the significant progress in AL outlined above, our study provides timely new insights into practical implementations, going well beyond the earlier survey by introducing new question groups on the relevance of data annotation and AL, as well as on trends in AL for NLP.

5 Discussion

We now first compare current perceptions to those from 2009 (RQ5)[7], then discuss the broader implications of our findings for AL as a field.

Comparison to Tomanek and Olsson (2009) The setup of AL has naturally shifted from maximum entropy approaches and support vector machines towards LLMs as the model of choice. Our survey indicates the popularity of smaller LLMs (in this study defined as models with less than 1B parameters). We hypothesize that the close or even on par performance these models can achieve with fine-tuning on many traditional supervised tasks (specifically encoders), combined with their lower hardware requirements, render larger models (i.e., LLMs with 1B parameters or more) computationally inefficient for practical application in AL. Larger LLMs may become the preferred choice in cases where a potentially minor performance increase is mission critical, or when no suitable smaller model exists (e.g., for languages where no strong encoder models are available). At the same time, uncertainty sampling remains the preferred

[7]The available data collection methodology does not allow us to claim representativeness of the NLP community for the two distinct samples. We therefore compare central findings rather than conducting longitudinal statistical analyses.

query strategy. Although surprising at first glance, given the growing range of more sophisticated alternatives, this is in accordance with the literature on strong performance of uncertainty sampling (Shen et al., 2017; Margatina et al., 2022; Schröder et al., 2022). In terms of stopping criteria, practical constraints such as time and budget still largely outweigh information-theoretical grounds.

Revisiting the central challenges identified by Tomanek and Olsson (2009) (cf. Section 1), our current survey reveals that many of these issues remain relevant. Skepticism about the practical effectiveness of AL and the perceived overhead of implementing annotation interfaces continue to prevent community members from adopting AL. Similarly, users of AL continue to report difficulties related to sampling complexity and interface design. However, we did not observe notable challenges related to annotator waiting times, unlike in 2009; on the contrary, participants indicated a willingness to tolerate considerable delays. The shift to LLMs may have altered requirements here.

Answer to RQ5: LLMs have substantially transformed AL by serving as powerful backbone models. Nonetheless, key challenges to AL adoption persist despite considerable growth in AL research.

Bridging the gap between research achievements and practical needs While LLMs offer options beyond serving as backbone models, these have not yet been integrated into the AL workflows reported by our participants. Approaches like prompt-based strategies (Li et al., 2024) and in-context learning (Margatina et al., 2023) could make AL more accessible for less-expert users, and using AL for few-shot example selection (Bayer and Reuter, 2024) may further reduce annotation effort. Trends such as LLMs-as-annotators, however, require stronger theoretical foundations first before they can provide reliable support in practice (Baumann et al., 2025; Tan et al., 2024).

Despite progress on single issues, such as model-dataset dependency (Tsvigun et al., 2022c; Jelenić et al., 2023) and performance stability (Li et al., 2024), research overall did not manage to sufficiently address the three overarching challenges to practical adoption that the community has raised:

(1) *Excessive complexity of the AL setup*: Respondents who never had applied AL most often cite a lack of methodological expertise and implementation overhead as obstructing factors. AL users share the concerns regarding methodological choices. The ongoing introduction of new algorithms (e.g., query strategies) and, thus, possible setup combinations continually increases complexity.

(2) *Risk of starting an AL project*: Respondents noted that the cost reduction through AL cannot be reliably estimated upfront. Its dependence on numerous variables makes a priori chosen setups prone to unsatisfactory outcomes (Lowell et al., 2019), possibly worse than random sampling (Ghose and Nguyen, 2024).

(3) *Perceived insufficiency of tooling*: Despite considerable activity in the development of AL annotation tools, their progress is still perceived as insufficient. We suspect that this is partly due to the large problem space, making it difficult for tools to serve all tasks.

As a starting point for addressing these challenges, we propose to:

(1) *prioritize the reduction of existing complexity instead of devising new algorithms*, e.g., by refuting the effectiveness of established strategies;

(2) *develop and continually research guidelines or heuristics for choosing AL components based on the setup's variables*, e.g., with a decision tree that specifies based on task and number of classes. More insights from practical use cases would also be immensely helpful to provide valuable points of comparison; and

(3) *enhance the visibility of existing tools, and provide user feedback* that can be used by tool developers to refine their interfaces to ease the application of AL (Shnarch et al., 2022b; Jukić et al., 2023), especially for non-expert users (Ras et al., 2022).

6 Conclusions

To gain comprehensive insights into AL in the era of LLMs, we conducted an online survey in the NLP community. Our findings suggest that AL is mostly viewed as a viable solution to the persistent challenge of data annotation. While recent research advancements have only partly permeated to AL applications yet, and the use of LLMs in AL is still on the rise, respondents are overall positive on the potential of LLMs in AL. We conclude with recommendations to improve three longstanding problems that hinder broader and more satisfying adoption, thereby hoping to strengthen AL further.

Limitations

Representativeness of the community survey Our results are descriptive for a sub-sample of the NLP community only, and therefore we cannot draw conclusions for the entirety of the community. As shown in Figure 3, our distribution strategy led to a potential over-representation of academia as a work sector and European countries (primarily Germany). While the geographic reach is broad (covering 44 countries in total), a stronger representation from North America and Asia was expected.[8] Moreover, as our survey is distributed in English, this may introduce a language bias, possibly excluding respondents who are not proficient in English or decide against responding due to being uncomfortable responding in English. However, distributing the survey in multiple languages seems infeasible given the vast number of written languages.

Temporal bias in perception It is difficult to determine when participants formed their opinions on AL, as assessments are commonly based on past experiences that may not align with current advancements. Similarly, we cannot ensure that recent improvements in AL were acknowledged or influenced their opinion. As a result, the collected responses may be subject to a temporal bias, potentially pointing to issues that have since been improved upon or even resolved. However, our analysis of recent projects confirms that the three key challenges (setup complexity, cost reduction estimation, and tooling) persist, to a certain extent.

Constraints of active learning in low-resource language scenarios Naturally, AL can be suitable in scenarios in which *annotated resources are scarce*, and yet *non-annotated data is available*. However, limitations apply where using machine learning-based methods to tackle NLP tasks may be prevented by the almost complete lack of language-specific data (be it annotated or non-annotated) as well as non-existing NLP pipelines and tools. As some of our survey participants pointed out in the free-form text fields, non-machine learning techniques often have to be applied for those languages, and in consequence, AL cannot provide a solution here. We recognize that especially in the low-resource language field non-machine learning based approaches are still relevant today.

[8] http://stats.aclrollingreview.org/submissions/geo-diversity/, https://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/

Ethical Considerations

To comply with data protection laws, we took the necessary steps to ensure compliance with the EU General Data Protection Regulation. Respondents granted their consent to the storage and analysis of their provided data, and the distribution of an anonymized version thereof, before starting the survey. To allow the NLP community to conduct extended analyses on the collected information and to uphold scientific transparency, we will publish the resulting dataset in a form that does not allow to infer conclusions about individuals.

Given the growing skepticism about data usage, particularly in the context of generative AI, and the voluntary nature of survey participation, we opted for a non-commercial license (i.e., CC BY-NC-SA 4.0[9]). This approach ensures that the data can be accessed and used for research purposes while safeguarding respondents' privacy and addressing concerns about potential misuse.

Acknowledgments

We thank all the great anonymous participants of the survey for their time and invaluable answers, which enabled this research. Likewise, we are grateful for everyone who supported us by distributing the survey, in particular Rabiraj Bandyopadhyay, David Berenstein, Ariel Gera, Alon Halfon, Philip H., Erik Körner, Gebriella Lapesa, Maximilian Maurer, Ines Montani, Paul Noirel, Julian Oestreich, Myrthe Reuver, Arij Riabi, Stefan Schweter, Artem Shelmanov, Eyal Shnarch, Manuel Tonneau and Akim Tsvigun. We are equally grateful to all anonymous supporters. Thanks to all tool authors and friends from the AL community for their support. We sincerely appreciate the contributions of all who helped, even if not mentioned by name.

We thank Ranjit Singh of *GESIS Survey Methods Consulting* for his expertise and feedback on early drafts of the survey, Harry Scells and Erik Körner who reviewed and tested the survey, Gerhard Heyer and Ingolf Römer for supporting the GDPR agreements, and Bildungsportal Sachsen for providing the LimeSurvey infrastructure, Constantin Hammer, Luisa Köhler, and Jens Syckor for supporting us with the GDPR-compliance in their roles as data protection officers, and Ana Lisboa Cotovio for support in the data collection for Figure 1.

[9] https://creativecommons.org/licenses/by-nc-sa/4.0/

Parts of this paper have been financially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI. Further parts of this work have been conducted within the CORAL project which was funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) under the grant number 16IS24077A. Responsibility for the content of this publication lies with the authors.

References

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *Preprint*, arXiv:2509.08825.

Markus Bayer and Christian Reuter. 2024. Activellm: Large language model-based active learning for textual few-shot scenarios. *Preprint*, arXiv:2405.10808.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Borisova, Raia Abu Ahmad, Leyla Garcia-Castro, Ricardo Usbeck, and Georg Rehm. 2024. Surveying the FAIRness of annotation tools: Difficult to find, difficult to reuse. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 29–45, St. Julians, Malta. Association for Computational Linguistics.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Daria Galimzianova and Leonid Sanochkin. 2024. Efficient active learning with adapters. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14374–14383, Miami, Florida, USA. Association for Computational Linguistics.

Abhishek Ghose and Emma Thuong Nguyen. 2024. On the fragility of active learners for text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22217–22233, Miami, Florida, USA. Association for Computational Linguistics.

Bertram Højer, Terne Sasha Thorn Jakobsen, Anna Rogers, and Stefan Heinrich. 2025. Research community perspectives on "intelligence" and large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25796–25812, Vienna, Austria. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Fran Jelenić, Josip Jukić, Nina Drobac, and Jan Snajder. 2023. On dataset transferability in active learning for transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2282–2295, Toronto, Canada. Association for Computational Linguistics.

Josip Jukić, Fran Jelenić, Miroslav Bićanić, and Jan Snajder. 2023. ALANNO: An active learning annotation system for mortals. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 228–235, Dubrovnik, Croatia. Association for Computational Linguistics.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X*, page 397–412, Berlin, Heidelberg. Springer-Verlag.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Philipp Kohl, Yoka Krämer, Claudia Fohry, and Bodo Kraft. 2024. Scoping review of active learning strategies and their evaluation environments for entity recognition tasks. In *Deep Learning Theory and Applications*, pages 84–106, Cham. Springer Nature Switzerland.

Dongseop Kwon, Sun Kim, Soo Yong Shin, and John Wilbur. 2013. Bioqrator: a web-based interactive biomedical literature curating system. In *Proceedings of the fourth biocreative challenge evaluation workshop*, volume 1, pages 241–246.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. Springer, ACM/Springer.

Dongyuan Li, Ying Zhang, Zhen Wang, Shiyin Tan, Satoshi Kosugi, and Manabu Okumura. 2024. Active learning for abstractive text summarization via LLM-determined curriculum and certainty gain maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8959–8971, Miami, Florida, USA. Association for Computational Linguistics.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024. Actively learn from LLMs with uncertainty propagation for generalized category discovery. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7845–7858, Mexico City, Mexico. Association for Computational Linguistics.

Mingnan Liu and Laura Wronski. 2018. Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, 36(1):116–124.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, Hong Kong, China. Association for Computational Linguistics.

Haocheng Luo, Wei Tan, Ngoc Nguyen, and Lan Du. 2023. Re-weighting tokens: A simple and effective active learning strategy for named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12725–12734, Singapore. Association for Computational Linguistics.

Katerina Margatina and Nikolaos Aletras. 2023. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419, Toronto, Canada. Association for Computational Linguistics.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vânia Mendonça, Ricardo Rei, Luísa Coheur, and Alberto Sardinha. 2023. Onception: Active learning with expert advice for real world machine translation. *Computational Linguistics*, 49(2):325–372.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2023. What do NLP researchers believe? results of the NLP community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada. Association for Computational Linguistics.

Charlotte Nachtegael, Jacopo De Stefani, and Tom Lenaerts. 2023. ALAMBIC : Active learning automation methods to battle inefficient curation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 117–127, Dubrovnik, Croatia. Association for Computational Linguistics.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. SICS Technical Report T2009:06, Swedish Institute of Computer Science.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. 2022. Explainable deep learning: A field guide for the uninitiated. *J. Artif. Int. Res.*, 73.

Anna Rogers and Sasha Luccioni. 2024. Position: Key claims in LLM research have a long tail of footnotes. In *Forty-first International Conference on Machine Learning*.

Julia Romberg and Tobias Escher. 2022. Automated topic categorisation of citizens' contributions: Reducing manual labelling efforts through active learning.

In *Electronic Government*, pages 369–385, Cham. Springer International Publishing.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *Preprint*, arXiv:2008.07267.

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V. Dylov. 2019. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, and Dakuo Wang. 2022a. Label sleuth: From unlabeled text to a classifier in a few hours. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 159–168, Abu Dhabi, UAE. Association for Computational Linguistics.

Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, Dakuo Wang, Lucy Yip, Liat Ein-Dor, Lena Dankin, Ilya Shnayderman, Ranit Aharonov, Yunyao Li, Naftali Liberman, Philip Levin Slesarev, Gwilym Newton, and 3 others. 2022b. Label sleuth: From unlabeled text to a classifier in a few hours. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 159–168, Abu Dhabi, UAE. Association for Computational Linguistics.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. 2023. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.

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

Katrin Tomanek and Fredrik Olsson. 2009. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48, Boulder, Colorado. Association for Computational Linguistics.

Manuel Tonneau, Dhaval Adjodah, Joao Palotti, Nir Grinberg, and Samuel Fraiberger. 2022. Multilingual detection of personal employment status on Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6564–6587, Dublin, Ireland. Association for Computational Linguistics.

Akim Tsvigun, Ivan Lysenko, Danila Sedashov, Ivan Lazichny, Eldar Damirov, Vladimir Karlov, Artemy Belousov, Leonid Sanochkin, Maxim Panov, Alexander Panchenko, Mikhail Burtsev, and Artem Shelmanov. 2022a. Active learning for abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5128–5152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Olga Shahmatova, Dmitry V. Dylov, Igor Galitskiy, and Artem Shelmanov. 2022b. ALToolbox: A set of tools for active learning annotation of natural language texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 406–434, Abu Dhabi, UAE. Association for Computational Linguistics.

Akim Tsvigun, Artem Shelmanov, Gleb Kuzmin, Leonid Sanochkin, Daniil Larionov, Gleb Gusev, Manvel Avetisian, and Leonid Zhukov. 2022c. Towards computationally feasible deep active learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1198–1218, Seattle, United States. Association for Computational Linguistics.

Michiel Van Der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Leon Weber and Barbara Plank. 2023. ActiveAED: A human in the loop improves annotation error detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8834–8845, Toronto, Canada. Association for Computational Linguistics.

Max Wiechmann, Seid Muhie Yimam, and Chris Biemann. 2021. ActiveAnno: General-purpose document-level annotation tool with active learning integration. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 99–105, Online. Association for Computational Linguistics.

Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. Hallucination diversity-aware active learning for text summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, and 15 others. 2025. From selection to generation: A survey of LLM-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, Vienna, Austria. Association for Computational Linguistics.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Kamer Yüksel, Ahmet Gunduz, Mohamed Al-badrashiny, and Hassan Sawaf. 2023. EvolveMT: an ensemble MT engine improving itself with usage only. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 341–346, Toronto, Canada. Association for Computational Linguistics.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Empirical evaluation of active learning techniques for neural MT. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, Hong Kong, China. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Appendix

A Active Learning Publications

To obtain the data for Figure 1, we start from the full ACL anthology in BibTeX format, including abstracts (January 14th, 2026). In a first step, only publications published in the context of AACL, ACL, COLING, EACL, EMNLP, NAACL, CL, and TACL between 2009 and 2025 were kept. This includes affiliated workshops. The resulting number of papers constitutes the *total number of publications*. Subsequently, we filtered publications that contain "active" in title or abstract. If "active" is directly followed by "learning", the publication is kept without further checks, in case of other next words, we manually checked the publication's relevance.[10] The resulting number of papers constitutes the *total number of AL publications*.

The numbers for 2025 must be interpreted with caution, as the ACL Anthology had not yet been fully updated with all entries at the time of analysis (e.g., the proceedings of AACL, some workshops).

B Survey Implementation

B.1 Survey Distribution and Advertisement

We chose a diverse distribution strategy, covering mailing lists, direct contacts, social media, and service providers. While some overlap between these channels may exist, we utilize all of them to ensure a wide outreach.

Mailing lists The **ACL mailing list** covers all current members of the Association of Computational Linguistics – about 10,400 as of end of 2024. *The worldwide recipients are researchers from academia and industry likewise.*[11] The **ELRA corpora-list** is managed by the European Language Resources Association. It serves as a platform for sharing information related to linguistic resources, therefore *providing a way to reach researchers and practitioners involved in data annotation in their daily work*. The exact number of recipients on the mailing list is not public; however, it is widely assumed as reaching a substantial audience. The members of **ELRA SIGUL-list** are professionals involved in the development of language resources and technologies for under-resourced languages. By utilizing this mailing list, we aim to *gain a more diverse perspective on supervised learning and AL for lesser-studied languages*, trying to mitigate potential biases in data collection driven by the uneven availability of language resources. The exact number of recipients on the mailing list is not public. **Natural Language Processing Data Community** is a google group[12] for *anyone interested in NLP including computational linguists, data scientists, and software engineers*. It counts 341 members as of the end of 2024. The **tada.cool initiative**[13] is a community of *interdisciplinary researchers, with a focus on machine learning and NLP for social science applications*. We reached 748 members via the slack channel.

Social media To gather more applied perspectives, we identified several chat communities on NLP and machine learning topics (e.g., PyTorch Lightning, Hugging Face Discord, and Hugging Face Posts) and sought approval to share the survey on these platforms. Approval was granted for the latter two, and the call for participation was subsequently posted. We also posted participation requests on several social media platforms (LinkedIn, X, and Bluesky), which are frequented by both audiences. These posts were re-shared by non-author accounts, further extending the survey's reach.

Researchers working on AL in NLP We curated a list of researchers who co-authored one or more publications on AL in major NLP venues over the past 15 years. To compile this list of experts, we started from the AL publications identified as described in Appendix A. We manually extracted the email addresses from publications that mentioned AL directly in the title. The approach also includes papers from workshops to capture more practical application experiences. We contacted all 601 identified individuals by a personalized mail. In 139 cases, the extracted e-mail addresses were no longer functional. We added 9 further personal contacts to complement the list.

Annotation tool providers As an additional strategy to incorporate practical insights from individuals involved in dataset creation efforts—potentially supported by AL—we reached out to individuals and companies responsible for annotation tools with AL solutions for text annotation.

To identify relevant tools, we supplemented a recent survey of annotation tools (Borisova et al., 2024) with additional tools we encountered through years of research in AL. We contacted: ActiveAnno (Wiechmann et al., 2021), AL Toolbox (Tsvigun et al., 2022b), Argilla[14], AWS Sagemaker Ground Truth and Comprehend, BioQRator (Kwon et al., 2013), CleanLab[15], INCEpTION (Klie et al., 2018), Labelbox[16], Label Sleuth (Shnarch et al., 2022a), Label Studio[17], MITRE

[10]This way, we account for wording variations such as "active fine-tuning" or "active sampling."

[11]The call to the ACL mailing list was sent out only once, due to two reasons: first, a manual inspection period of about two weeks by the administrators before distribution limited us to requesting only two mailings during the survey period; second, our second request was deleted for unclear reasons.

[12]https://groups.google.com/a/datacommunitydc.org/g/nlp

[13]https://sites.google.com/view/polsci-ml-initiative/talks

[14]https://github.com/argilla-io/argilla

[15]https://github.com/cleanlab/cleanlab

[16]https://labelbox.com/

[17]https://labelstud.io/

Annotation Toolkit (MAT)[18], and POTATO (Pei et al., 2022). We also contacted several companies and services via their official communication channels, asking them to forward the survey to their users. As an incentive, potential insights that could be gained into the actual needs of users in supervised learning and AL were emphasized. Although this approach yielded limited feedback, it nevertheless contributed to the overall outreach. Positive responses were received from AL Toolbox, which forwarded the call to some of their annotators; Argilla, which permitted the call to be shared in the Hugging Face Discord channel; Explosion, which shared the call on LinkedIn; Labelbox, which forwarded the call internally; and Label Sleuth, which forwarded the call to their users.

In retrospect, we believe the limited engagement from tool providers and companies may be attributed to a combination of factors: an insufficiently compelling appeal or reward for sharing that we could provide from our side, the possibility that the survey was not a suitable fit for their interests, and the timing of the survey launch.

B.2 Survey Portal & Execution

The survey was conducted using LimeSurvey in version 3.27.4. Participants could revise their answers at any time; they could navigate back and forth, as long as the survey had not been completed.

B.3 Participation and Completion Rates

The survey remained open for six weeks. According to question VII.7, mailing lists were the most effective referral source, with noticeable spikes in activity following each call (with declining impact). Given this pattern, and after three waves of invitations, we did not expect a significant increase in responses by keeping the survey open longer.

Given that we have no information about the number of subscribers to the various mailing lists, or who we reached via social media or other means, it is impossible to determine how many individuals received the invitation to participate in the survey. According to LimeSurvey statistics, 1124 individuals clicked the survey link. Of these, 171 began the main content of the survey, and 144 completed it.

B.4 Data Protection

In consultation with data protection officers, we implemented measures to ensure compliance with the EU General Data Protection Regulation. Following it's principle of data minimization, we collected only data necessary for our research objectives. The survey was conducted anonymously, with the option to provide an email address to receive the survey evaluation afterward. Free-form text fields are a gray area, as participants may potentially enter personally identifiable information, particularly when combined with demographic data collected. To meet legal obligations, we informed participants about how any personal data would be processed, and concluded a joint-controller agreement among the institutions involved.

B.5 Anonymization for Dataset Release

For public release, text answers were manually processed as follows. We started from the raw dataset, excluding the separately stored email addresses. For each field, we replaced all mentions that could reveal the respondents' identities with placeholders, including URLs, links, email addresses, single words, sentences, and paragraphs. We also applied this to language mentions, as they could contribute to de-anonymization when combined with other fields (e.g., in case of languages that few people in the NLP community are researching). In this context, we took a conservative approach to V.2 and V.7 by removing all answers to preserve anonymity. For question items containing highly specific information that may allow to draw conclusions about the respective participant, we replaced the entries with broader categories: **IV.1 (year):** 2020–2024, 2015–2019, 2010–2014, 2005–2009; **V.3 (hours):** up to 20, 21–50, 51–100, 101–500, more than 500; **V.4 (instance count):** up to 100, 101–500, 501–1000, 1001–10,000, more than 10,000, N/A; **V.4 (instance type):** documents, sentences, tokens, other, N/A; **VII.5 (countries):** Africa, Asia, Europe, North America, Oceania, South America.

Responses and demographics are released separately and randomly shuffled to prevent re-merging.

C Full Survey

Data Annotation Bottleneck & Active Learning for NLP in the Era of LLMs

The success of Natural Language Processing (NLP) often depends on the availability of (high-quality) data. In particular, **the costly manual annotation of text data has posed a major challenge since the early days of NLP**. To overcome the data annotation bottleneck, a number of methods have been proposed. One prominent method in this context is **Active Learning**, which aims to minimize the set of data that needs to be annotated.

[18]https://sourceforge.net/projects/mat-annotation/

However, the development of Large Language Models (LLMs) has changed the field of NLP considerably. For this reason, it is of huge interest to us working in this field (both in research and in practical application) to understand **if and how a lack of annotated data is still affecting NLP today**.

At the center of this survey is Active Learning, which was last surveyed in a web survey in 2009. Fifteen years later, we aim to reassess the current state of the method from the user's point of view. Besides inquiring where Active Learning is used, we also ask where it is not used in favor of other methods. Moreover, we want to understand which computational methods the community considers most useful to overcome a lack of annotated data.

The survey is conducted solely for non-commercial, academic purposes. It specifically targets participants who are or have been involved in supervised machine learning for NLP. Knowledge about Active Learning is not required. Filling out the survey will take you approximately 15 minutes.

Why should I invest my time in this survey? We need your collective expertise in the field of NLP, Supervised Machine Learning, or Active Learning, to understand how recent advancements, such as LLMs, have changed the long-standing data annotation bottleneck. The results of this survey will help the community to better understand the state and open issues of contemporary Active Learning, and incorporate these insights into research and development of new methods and technologies. To this end, a study presenting and discussing the results of this survey will be published as an open access publication. If you wish to be notified upon publication, you can optionally enter your email address at the end of the survey.

What is Active Learning? Active Learning is a method to create a small but meaningful annotated dataset with the goal of minimizing the annotation effort. It is an iterative cyclic method between a human annotator and a learning algorithm. In each iteration, an instance selection strategy (also referred to as query or acquisition strategy) is used to select data points that are considered particularly useful to be annotated next. These can be, for example (among many other strategies), the instances for which a machine learning model is most uncertain.

The survey is initiated by the researchers *[anonymized for review]*. If you have any questions, please contact us via *[anonymized for review]*.

Consent, Data Usage, and Privacy

Declaration of Consent
[not disclosed to ensure anonymity in the review]

We need your consent before we can start collecting data for the survey.

() Yes, I would like to participate in this survey.

() No, I would not like to participate in this survey.

Data Usage and Privacy Information: We will store your provided answers on a server located in *[anonymized for review]*. After the survey has been conducted, we will process the collected data in order to investigate the research questions presented above to investigate if and how a lack of annotated data still affects the field of NLP in 2024. Our goal is to make the collected data available to the community under a non-commercial share-alike CC BY-NC-SA 4.0 license. ***The public dataset will be completely anonymous.*** Any information in free text responses that could identify specific respondents will be removed before publication of the dataset. The processing of any personal data is detailed in privacy information in compliance with GDPR. This includes information on the purpose of data collection, how long we retain your data, your rights regarding your data, and how to contact us with any concerns.

We need your confirm that you have read and consented to the data usage and privacy information.

() I have read the information on data usage and privacy information and consent that my answers given in this survey will be stored and analyzed for research purposes as described above. I agree that my contribution will be made available to the community as part of a public dataset under CC BY-NC-SA 4.0 license. I hereby grant the authors permission to distribute my data under these terms. It will be ensured that no conclusions about individuals can be inferred from this data.

I. Relevance of Annotated Data in Light of Recent Advancements in NLP

This survey focuses on supervised learning in NLP, where annotated data is used to train machine learning models.

1. [sc,m][19] When applying NLP methods, have you ever encountered difficulties due to a lack of annotated data (e.g., you had to build a new dataset, without which the desired goal was unattainable)?
   (a) Yes
   (b) No *[directs participants to part VII.]*

You stated that you have had difficulties due to a lack of data.

2. [mc,m] Under what circumstances did you encounter difficulties due to lack of annotated data? Select all that apply.
   (a) When working on following under-resourced languages: *[text input]*
   (b) When working on a certain task (e.g., summarization): *[text input]*
   (c) When working with a specific requirement (e.g., on a problem with many classes): *[text input]*
   (d) Other: *[text input]*

Recent advancements, such as LLMs and Generative AI, have profoundly changed the NLP landscape. Considering these developments, do you think that a lack of annotated data is **currently** a problem? Please indicate below, how much you agree or disagree with each of these statements.

3. [sc,m] The type of tasks that I used to solve with supervised learning, can be (or will soon be) solved by leveraging generative AI to synthesize training data.
   - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*

4. [sc,m] Many problems will still be solved as supervised learning tasks.
   - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*

[19]Question properties: single choice (sc), multiple choice (mc), mandatory (m), and display conditional on previous response (c).

5. [sc,m] Annotated data is still a limiting factor in the context of supervised learning.
   - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*
6. [sc,m] Annotated data is still a limiting factor for the languages I work with.
   - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*
7. [sc,m] Annotated data is still a limiting factor for problems of a certain complexity (e.g., in summarization where summarizing long documents can be time-consuming).
   - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*

II. Overcoming a Lack of Annotated Data

Annotated data is required for supervised learning. In this section, we ask about methods to overcome a lack of data in such situations, and about your experiences with them.

1. [sc,m] Have you ever used a computational method to overcome a lack of annotated data?
   (a) Yes
   (b) No
2. [mc,m,c:II1] Please choose all computational methods that you have ever used to overcome a lack of annotated data.
   (a) Active Learning (e.g., to efficiently annotate data)
   (b) Data augmentation (e.g., to use existing training instances more efficiently)
   (c) Few-shot Learning (e.g., to achieve competitive results with only a few instances)
   (d) Semi-supervised Learning (e.g., use already annotated data for automatic labeling of training instances)
   (e) Text generation (e.g., to generate training instances)
   (f) Transfer learning (e.g., to use the pre-existing knowledge of already trained models)
   (g) Weak supervision (e.g., to programmatically obtain (pseudo-)annotated instances using, e.g., labeling functions)
   (h) Zero-shot (e.g., to work without any training data at all)
   (i) Other: *[text input]*
3. [mc,m,c:II1] What are the reasons why you have never used a computational method to overcome a lack of annotated data? Please choose all options that apply to you.
   (a) I have never heard of these methods.
   (b) I never coordinated an annotation project.
   (c) We had enough human resources to annotate our data.

You stated that you have used a method other than Active Learning.

4. [sc,m,c:II2] Did you consider your chosen computational method(s) to overcome a lack of annotated data to be successful?
   - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*
5. [mc,m,c:II2] What are the specific reasons for not applying Active Learning? Please select all that apply.
   (a) Never heard of Active Learning.
   (b) Insufficient expertise/knowledge.
   (c) Did not want to spend overhead in implementing an Active Learning-based annotation environment.
   (d) Did not know of any suitable annotation tools that easily integrate Active Learning.
   (e) Did not meet my project's specific requirements.
   (f) Wanted to avoid sampling bias in the corpus.
   (g) Was not convinced that Active Learning would reduce annotation cost.
   (h) Did not use Active Learning because I couldn't estimate upfront its impact on reducing annotation costs.
   (i) Other: *[text input]*
6. [m,c:II5] Why did Active Learning not fit your specific requirements? *[text input]*
7. [sc,m,c:II2] Would you consider applying Active Learning in future annotation projects?
   (a) Yes
   (b) No
   (c) I don't know enough about Active Learning to answer this question.
   (d) Only under specific circumstances: *[text input]*
8. [mc,m,c:II7] What are the reasons why you would not apply Active Learning in future annotation projects?
   (a) Active Learning has become obsolete nowadays.
   (b) Active Learning is a useful concept in theory but does not work well enough in practice.
   (c) Other: *[text input]*

[directs participants who never used Active Learning to part VII.]

III. Active Learning – General

In this section, we want to learn more about your experience with Active Learning.

1. [mc,m] What is your expertise with Active Learning?
   (a) I have been part of an Active Learning workflow as an annotator.
   (b) I have organized an Active Learning annotation workflow.
   (c) I have knowledge about instance selection strategies in Active Learning but use pre-built solutions.
   (d) I have implemented Active Learning strategies or workflows.
   (e) I have researched Active Learning.
2. [mc,m] What was your primary motivation to use Active Learning?
   (a) Obtain annotated data at minimal annotation costs
   (b) Gather experience with practical applications of Active Learning
   (c) Test your annotation guidelines to identify difficult examples

(d) Other: *[text input]*

In the following questions, we sometimes ask about *(annotation) instances*, by which we mean the entities to be annotated, such as documents, words, or sentences.

3. [mc,m] What obstacles have you encountered when applying Active Learning?
   (a) Finding annotators that achieve satisfactory performance
   (b) Performing Active Learning with multiple annotators (e.g., distributing work, resolving disagreement)
   (c) Unclear instructions how to annotate (lack of annotation guidelines)
   (d) Hard cases for annotation (undecidable, ambiguous, subjective)
   (e) Lack of resources (GPU/TPU)
   (f) Choice of model and instance selection algorithm
   (g) Choice of stopping criterion
   (h) Other: *[text input]*

4. [sc,m] Between two iterations of Active Learning, a new model is usually trained. This new model is then used to select new examples for annotation. What is the maximum time you would be willing to wait between two annotation cycles?
   (a) Less than a minute
   (b) 1 minute to less than 10 minutes
   (c) 10 minutes to less than 1 hour
   (d) 1 hour to less than 10 hours
   (e) 10 hours to less than 1 day
   (f) 1 day to less than 1 week
   (g) I will wait long as it takes

5. [mc,m] In which scenario(s) have you worked with Active Learning?
   (a) In a lab setting with simulated annotators
   (b) In an applied setting with human annotators

[directs participants that never used Active Learning in an applied setting to part VI.]

IV. Active Learning – Methodological Setup

The following questions relate to type and scope of a previous annotation project in which you employed Active Learning.

In case you performed more than one such annotation project, please think of your **most recent project** with Active Learning when answering the following questions.

1. [m] In which year did the project start? *[text input]*

2. [sc,m] Which type of machine learning model was trained during Active Learning?
   (a) Classic (Naive Bayes, Support Vector Machine, Conditional Random Fields, ...)
   (b) Neural Network (Convolutional Neural Network, Long Short-term Memory, ...)
   (c) BERT era Language Models such as BERT, GPT-2, RoBERTa, ...
   (d) Large Language Models such as Phi-3, Llama, Mistral, Gemma, Gemini, GPT-3/4, ...
   (e) Do not know
   (f) Other: *[text input]*

3. [sc,m] What instance selection strategy[20] have you used during Active Learning? If you have tried different strategies, please indicate the strategy that you eventually decided on.
   (a) Uncertainty-based[21]
   (b) Disagreement-based[22]
   (c) Gradient information-based[23]
   (d) Performance prediction-based[24]
   (e) Density-based[25]
   (f) Discriminative-based[26]
   (g) Batch diversity[27]
   (h) Hybrid (Combines different of the aforementioned concepts.)
   (i) Do not know
   (j) Other: *[text input]*

4. [sc,m] Between two iterations of Active Learning, a new model is usually trained. This new model is then used to select new examples for annotation. How long was the resulting waiting time for the annotator in your scenario on average? Please select the matching time interval.
   (a) We did not retrain
   (b) Less than a minute
   (c) 1 minute to less than 10 minutes
   (d) 10 minutes to less than 1 hour
   (e) 1 hour to less than 10 hours
   (f) 10 hours to less than 1 day
   (g) 1 day to less than 1 week
   (h) Do not know

5. [sc,m] How did you decide when to stop the Active Learning process?
   (a) Evaluation of the learned model on a held-out gold standard
   (b) Money and/or time available for annotation were depleted
   (c) All relevant documents annotated

[20] Tooltip: The most salient component of an Active Learning setup is the instance selection strategy, which decides on the instances to be labeled next.

[21] Tooltip: Selects instances for which the model is most uncertain. Examples are least-confidence, entropy and margin-sampling.

[22] Tooltip: Uses multiple models to select instances based on disagreement between model outputs. A popular example is query-by-committee.

[23] Tooltip: Selects instances based on their impact on the model. The impact is measured by, e.g., the norm of the gradients. A popular example is expected gradient length.

[24] Tooltip: Selects instances that have the most potential of reducing future errors. Examples are policy-learning strategies with reinforcement learning or imitation learning, and cartography Active Learning.

[25] Tooltip: Selects instances that are representative of dense regions in the embedding space. Density-based representatives can be selected, e.g., based on n-gram counts.

[26] Tooltip: Selects instances that differ from already annotated instances. Examples are selecting discriminative-based representatives based on rare words or a lesser similarity to the already annotated instances, as well as the use of discriminative Active Learning.

[27] Tooltip: Selects a batch of instances that are diverse. Examples are coresets, BADGE and ALPS.

(d) The stopping time was indicated by an algorithmic stopping criterion (for example based on the current model's performance)
(e) Do not know
(f) Other: *[text input]*

V. Active Learning – Task & Annotation Setup

The following questions relate to type and scope of a previous annotation project of your choice, where Active Learning was employed.

In case you performed more than one such annotation project, please think of your **most recent project** with Active Learning when answering the following questions.

1. [sc,m] What specific NLP task did you collect annotations for?
   (a) Automatic speech recognition
   (b) Coreference resolution
   (c) Chunking
   (d) Information extraction
   (e) Language generation
   (f) Language understanding
   (g) Machine translation
   (h) Morphological analysis
   (i) Named entity recognition
   (j) Part-of-speech tagging
   (k) Question answering
   (l) Relation extraction
   (m) Semantic similarity
   (n) Sentiment analysis
   (o) Syntactic parsing
   (p) Summarization
   (q) Text categorization
   (r) Word segmentation
   (s) Word sense disambiguation
   (t) Other: *[text input]*

2. [mc,m] What was the language of the texts that were annotated? (If multilingual, select multiple options.)
   (a) Arabic
   (b) English
   (c) French
   (d) German
   (e) Hindi
   (f) Mandarin
   (g) Spanish
   (h) Other: *[text input]*

3. [sc,m] How much time was spent on annotation in total? In the case of multiple annotators, please enter the total sum of hours worked by the annotators.
   (a) *[text input]* hours
   (b) Do not know

4. [m] What was the size of the resulting annotated corpus in terms of annotated instances? Please provide the annotation instance type in brackets, e.g. 1000 (tokens), 100 (sentences), or 10 (documents).. *[text input]*

5. [mc,m] Who were your annotators?
   (a) Annotation was done by the project coordinators or other project members (data scientist / linguist roles or similar)
   (b) Domain experts
   (c) Non-domain experts
   (d) Other: *[text input]*

6. [sc,m] Did you use an annotation tool with Active Learning support?
   (a) Yes
   (b) No

7. [sc,m,c:V6] Which annotation tool with Active Learning support did you use?
   (a) ActiveAnno
   (b) Argilla
   (c) AWS Sagemaker Ground Truth
   (d) AWS Comprehend
   (e) BioQRator
   (f) INCEpTION
   (g) Labelbox
   (h) Label Studio
   (i) MAT
   (j) Potato
   (k) Prodigy
   (l) Other: *[text input]*

8. [sc,m] In annotation, what did you do with data points that were ambiguous (i.e., caused disagreement between multiple annotators)?
   (a) Deleted them
   (b) Assigned one of two labels
   (c) Not applicable, each instance had at most one annotation.
   (d) Other: *[text input]*

9. [sc,m] While using Active Learning, did you have to change the annotation schema due to challenging examples surfaced during the annotation process (eg, add new classes in a classification task)?
   (a) No
   (b) Yes

10. [sc,m,c:V9] What did you do with previously annotated examples?
    (a) Previously annotated examples were re-annotated.
    (b) Previously annotated examples were left unchanged.
    (c) Other: *[text input]*

11. [sc,m] Did you consider your annotation project to be successful?
    - Very successful, Successful, Neither unsuccessful nor successful, Unsuccessful, Very unsuccessful *(5-point Likert scale)*

12. [sc,m] Was the use of Active Learning as effective as intended?
    - Very effective, Effective, Neither effective nor ineffective, Ineffective, Very Ineffective *(5-point Likert scale)*

13. [mc,m,c:V12] What were the reasons that reduced the effectiveness of Active Learning in your case of application? Please choose all options that apply.
    (a) In retrospect, I had insufficient expertise/knowledge.
    (b) Overhead in setting up an Active Learning-based annotation environment.
    (c) Lack of suitable annotation tools that easily integrate Active Learning.
    (d) In retrospect, Active Learning did not meet my project's specific requirements.
    (e) Active Learning created a sampling bias in the corpus.
    (f) Active Learning did not reduce annotation cost.
    (g) Active Learning did not work well in my scenario.
    (h) My upfront estimation of the impact on reducing annotation costs was not accurate.
    (i) The dependency of the model from the created dataset and vice versa.
    (j) Other: *[text input]*

14. [sc,m] Would you consider applying Active Learning again in future annotation projects?
    (a) No
    (b) Yes
    (c) Other: *[text input]*

15. [mc,c:V14] Why would you not consider Active Learning in future annotation projects?
    (a) Active Learning has become obsolete nowadays.
    (b) Active Learning is a useful concept in theory but does not work well enough in practice.
    (c) Other: *[text input]*

VI. Trends in Active Learning for NLP

In this section, we are interested in your opinion on whether the technological and methodological developments have affected the field of Active Learning, and what you think is still missing.

Please indicate how much you agree or disagree with the following statements:

1. [sc,m] Text embeddings are crucial for many methods in Active Learning for NLP.
    - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*
2. [sc,m] GPU-accelerated computing has changed the choice of model in Active Learning for NLP.
    - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*
3. [sc,m] Neural Networks have changed the choice of instance selection strategy in Active Learning for NLP.
    - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*
4. [sc,m] LLMs (with 1B parameters or more) are frequently used in Active Learning for NLP.
    - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree *(5-point Likert scale)*

5. What are obvious next developments for Active Learning in NLP that you believe will have a big impact on the field? *[text input]*

VII. Background Information

You are almost done. We just need a few more basic pieces of information about you in order to put your answers into context.

1. [mc] Where are you currently working? Please choose all options that apply to you.
    (a) Academia
    (b) Industry
    (c) Governmental organization
    (d) Other: *[text input]*

2. [mc] What is your education background? Please choose all options that apply to you.
    (a) Linguistics
    (b) Computational Linguistics
    (c) Computer Science/Informatics
    (d) Engineering
    (e) Mathematics
    (f) Social Sciences
    (g) Other: *[text input]*

3. [mc] Which field are you currently working in? Please choose all options that apply to you.
    (a) Economics
    (b) Education
    (c) Financial sector
    (d) Health services sector
    (e) Humanities
    (f) Information and communications technology
    (g) Legal sector
    (h) Public service sector
    (i) Society and politics
    (j) Other: *[text input]*

4. [sc] How many years of experience do you have in Machine Learning/NLP?
    (a) Up to 1 year
    (b) 1-2 years
    (c) 3-5 years
    (d) 5-10 years
    (e) More than 10 years

5. [sc] Which country is your primary place of residence?
    (a) *Dropdown menu with 196 country names*
    (b) Other: *[text input]*

6. Do you have any comments on the survey? *[text input]*

7. [sc] How did you become aware of the survey?
    (a) Social media (e.g., Facebook, Twitter, Instagram)
    (b) Personalized email invitation
    (c) Mailing list (e.g., ACL Member Portal)
    (d) Friend, colleague, or peer shared it
    (e) Other: *[text input]*

D Results

Table 3 provides a full overview of survey participants' responses to the predefined options for all 52 questions. It is beyond the scope of this paper to cover all analyses enabled by the collected data, but the public release will allow researchers to conduct in-depth analyses of further aspects.

D.1 Conservative Estimates (No Authors)

Our targeted recruitment of authors may have led to potential overestimation of AL in question groups I and II (the non-AL parts of the survey). To transparently account for potential author bias (response bias), our data includes information on recruitment methods (see VII.7). By excluding the 33 participants recruited via personalized mails, we can provide a *conservative baseline*. Excluding authors of AL publications provides the following picture:

RQ1 106 out of 111 participants had encountered a lack of annotated data. With respect to potential changes due to recent advancements in NLP, participants largely agree that many problems will still require supervised learning (81%; vs. 80% in the full participant sample), and that annotated data remains a limiting factor; in general (93%; vs. 94% in the full sample), for certain languages (75%; vs. 75% in the full sample) and for problems of a certain complexity (92%; vs. 91% in the full sample). Fewer participants agree that generative AI for training data synthesis offers a viable solution (30%; vs. 30% in the full sample). **The response distribution of the conservative baseline closely mirrors that of the full participant sample.**

RQ2 89 out of the 106 respondents had used computational methods.

Adoption of active learning: AL was used by 36 respondents (40%; vs. 53% in the full sample). When asking why participants had not chosen AL, 25% (vs. 25% in the full sample) had never heard of it, and 51% (vs. 49% in the full sample) cited insufficient expertise. Methodological concerns cover the expected implementation overhead (40%; vs. 37% in the full sample), unsuitable annotation tools (34%; vs. 32% in the full sample), the difficulty of estimating effectiveness upfront (21%; vs. 12% in the full sample), and sampling bias (15%; vs. 18% in the full sample). 9% (vs. 9% in the full sample) found it unsuitable due to specific project requirements, and 26% (vs. 14% in the full sample) doubted AL's effectiveness at all.

Despite these concerns, 54% (vs. 54% in the full sample) of those who had not yet used AL consider it for future projects and further 37% (vs. 38% in the full sample) indicate uncertainty due to a lack of knowledge. One participant would explicitly not use AL, four made its use dependent on the specific circumstances (similar to the full sample). **While response patterns are similar in general, the conservative sample emphasizes more the difficulty of upfront estimation of and general doubts about AL's effectiveness.**

Alternative methods: The current users of AL also report obstacles in data annotation with AL. These are: finding annotators (36%; vs. 35% in the full sample), AL with multiple annotators (53%; vs. 56% in the full sample), unclear annotation instructions (31%; vs. 29% in the full sample), hard cases for annotation (53%; vs. 59% in the full sample), lack of GPU/TPU resources (22%; vs. 25% in the full sample), choice of model / query strategy (44%; vs. 49% in the full sample), choice of stopping criterion (36%; vs. 35% in the full sample), and other (6%; vs. 6% in the full sample). **The response distributions closely resemble each other.**

The use of computational alternatives to AL is widely dispersed across participants: data augmentation (70%; vs. 73% in the full sample), few-shot learning (57%; vs. 63% in the full sample), semi-supervised learning (51%; vs. 53% in the full sample), text generation (46%; vs. 49% in the full sample), transfer learning (62% vs. 66% in the full sample), weak supervision (40%; vs. 44% in the full sample), zero-shot learning (48%; vs. 53% in the full sample), and other (12%; vs. 12% in the full sample). **In the conservative baseline sample, AL is among the least used methods.** The alternative methods were considered successful in overcoming a lack of annotated data in slightly over half of the cases (56%; vs. 57% in the full sample).

D.2 Sector-specific Analysis of Section 3.4

In Section 3.4, we assess the practical utility of contemporary AL by including respondents from academia, industry and governmental organizations. This way, we do not make a distinction between experiences reported from NLP researchers in academic applications and practical utility "in the real world", as we believe that using AL-supported annotation for the creation of research data also offers valid practical insights (in contrast to the AL-simulation lab scenario without human annotators). Nonetheless, there may be considerable differences in usage that should be analyzed further to inform future work. What follows, therefore, presents a distinction between academic (16) and industrial/governmental respondents (20, with 3 overlaps) based on the 33 recent projects.

Models, query strategies, and stopping criteria Both groups show similar preferences for LLMs,

VARIABLE	ACADEMIA			INDUSTRY/GOVERNMENT		
	# no	# yes	μyes	# no	# yes	μyes
Insufficient expertise (a)	9	2	0.18	14	1	0.07
Overhead in setup (b)	7	4	0.36	11	4	0.27
Lack of annotation tools (c)	6	5	0.45	9	6	0.4
Project-specific requirements (d)	10	1	0.09	14	1	0.07
Sampling bias (e)	8	3	0.27	12	3	0.20
No cost reduction (f)	11	0	0.00	12	3	0.20
Did not work well (g)	7	4	0.36	11	4	0.27
Inaccurate upfront estimation (h)	10	1	0.09	12	3	0.2
Dataset-model dependency (i)	7	4	0.36	13	2	0.13

Table 2: Predefined reason selection for reduced effectiveness in AL (V.13) in practical projects (2020 onward) across sectors. We show the frequency of no (# no; the respective reason did not affect the effectiveness of AL) and yes answers (# yes; the respective reason affected the effectiveness of AL), and the mean of yes answers (μ_{yes}).

with about two-thirds in each group using them, primarily the smaller models. Industry/governmental participants rely more on uncertainty sampling (50% vs. 25%), and their waiting times vary more—possibly due to a larger data pool.

Annotation tools for active learning Annotation tool use differs notably: 70% of industry/governmental participants use tools vs. only 23% of those that exclusively work in academia.

Satisfaction with project The answer distributions on the Likert scales for success and effectiveness differed slightly between industry/government and academia. However, the main trend is clear: 85% respectively 94% found their projects (very) successful, and 65% respectively 63% found them (very) effective. In case of the 23 cases of reduced effectiveness, Table 2 provides an overview of the response patterns for reasons. Overall, both groups exhibit similar tendencies in the yes/no responses. Especially, for the lack of suitable annotation tools (c), unmet project-specific requirements (d), and sampling bias (e), the mean values in the sample at hand are closely aligned. Slightly more difference can be observed in an inaccurate upfront estimation (h) or no reduction (f) of annotation costs, with no academic respondents indicating the latter issue. In contrast, model-dependency of datasets (i) seems to be more of a concern in academia, as is insufficient expertise (a).[28] 90% of industry/governmental respondents would use AL again without restrictions, compared to 81% of academic respondents, who are slightly more critical.

To sum up, while there are *some usage differences, both groups align closely on effectiveness and reasons for reduced effectiveness*, with only minor variations. Our findings support the identified main challenges for contemporary AL across NLP researchers and "real-world" applicants.

[28]We refrain from applying significance tests due to the small sample size, which limits the generalizability of the observations and, thus, the informative value of such analysis.

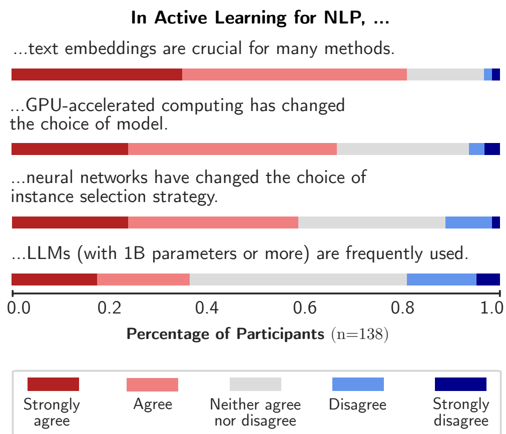


Figure 9: Participants' assessment of the impact of various methodological and technological developments on AL using a 5-point Likert scale (cf. VI.1–4).

D.3 Selected Additional Analysis

Order of annotated data The free-text answers to V.4 inform us about the volume of annotated data collected in the reported annotation projects. For the 33 recent applied AL projects, we observe a range from 50 to 1,000,000 annotation units, with an average of 43,265 (three entries miss size information). Examining differences between industry and academic respondents, we find that the former report creating substantially less annotated data (on average 2,313 units) than the latter (on average 86,506 units). This may reflect more restrictive requirements in industrial applications.

Impact of recent developments in NLP on AL Table 9 illustrates the perception of how recent developments in NLP have impacted AL.

Survey Question	Mapping to RQ	n	Responses (total / percentage of participants)
I.1	RQ1	144	Yes (138 / 96%), No (6 / 4%)
I.2	RQ1	138	When working on following under-resourced languages (77 / 56%), When working on a certain task (96 / 70%), When working with a specific requirement (71 / 51%), Other (119 / 86%)
I.3	RQ1	138	Strongly agree (3 / 2%), Agree (38 / 28%), Neither agree nor disagree (38 / 28%), Disagree (39 / 28%), Strongly disagree (20 / 14%)
I.4	RQ1	138	Strongly agree (35 / 25%), Agree (75 / 54%), Neither agree nor disagree (18 / 13%), Disagree (8 / 6%), Strongly disagree (2 / 1%)
I.5	RQ1	138	Strongly agree (59 / 43%), Agree (71 / 51%), Neither agree nor disagree (6 / 4%), Disagree (1 / 1%), Strongly disagree (1 / 1%)
I.6	RQ1	138	Strongly agree (51 / 37%), Agree (53 / 38%), Neither agree nor disagree (25 / 18%), Disagree (8 / 6%), Strongly disagree (1 / 1%)
I.7	RQ1	138	Strongly agree (62 / 45%), Agree (63 / 46%), Neither agree nor disagree (11 / 8%), Disagree (1 / 1%), Strongly disagree (1 / 1%)
II.1	RQ2	138	Yes (120 / 87%), No (18 / 13%)
II.2	RQ2	120	Active Learning (63 / 53%), Data augmentation (88 / 73%), Few-shot Learning (75 / 63%), Semi-supervised Learning (64 / 53%), Text generation (59 / 49%), Transfer learning (79 / 66%), Weak supervision (53 / 44%), Zero-shot (63 / 53%), Other (14 / 12%)
II.3*	RQ2	18	I have never heard of these methods. (10 / 56%), I never coordinated an annotation project. (6 / 33%), We had enough human resources to annotate our data (5 / 28%).
II.4	RQ2	56[29]	Strongly agree (4 / 7%), Agree (28 / 50%), Neither agree nor disagree (21 / 38%), Disagree (2 / 4%), Strongly disagree (1 / 2%)
II.5	RQ2, RQ5	57	Never heard of Active Learning. (14 / 25%), Insufficient expertise / knowledge. (28 / 49%), Did not want to spend overhead in implementing an Active Learning-based annotation environment. (21 / 37%), Did not know of any suitable annotation tools that easily integrate Active Learning. (18 / 32%), Did not meet my project's specific requirements. (5 / 9%), Wanted to avoid sampling bias in the corpus. (10 / 18%), Was not convinced that Active Learning would reduce annotation cost. (14 / 25%), Did not use Active Learning because I couldn't estimate upfront its impact on reducing annotation costs. (12 / 21%), Other (9 / 16%)
II.6*	RQ2, RQ5	5	*free text answer* (5 / 100%)
II.7	RQ2, RQ5	56	Yes (30 / 54%), No (1 / 2%), I don't know enough about Active Learning to answer this question. (21 / 38%), Only under specific circumstances (4 / 7%)
II.8	RQ2	1	Active Learning has become obsolete nowadays. (0 / 0%), Active Learning is a useful concept in theory but does not work well enough in practice. (1 / 100%), Other (0 / 0%)
III.1*	RQ2, RQ3	63	I have been part of an Active Learning workflow as an annotator. (24 / 38%), I have organized an Active Learning annotation workflow. (38 / 60%), I have knowledge about instance selection strategies in Active Learning but use pre-built solutions. (28 / 44%), I have implemented Active Learning strategies or workflows. (45 / 71%), I have researched Active Learning. (43 / 68%)
III.2	RQ2, RQ5	63	Obtain annotated data at minimal annotation costs (55 / 87%), Gather experience with practical applications of Active Learning (29 / 46%), Test your annotation guidelines to identify difficult examples (15 / 24%), Other (5 / 8%)
III.3	RQ2	63	Finding annotators that achieve satisfactory performance (22 / 35%), Performing Active Learning with multiple annotators (35 / 56%), Unclear instructions how to annotate (18 / 29%), Hard cases for annotation (37 / 59%), Lack of resources (16 / 25%), Choice of model and instance selection algorithm (31 / 49%), Choice of stopping criterion (22 / 35%), Other (4 / 6%)
III.4**	RQ3	63	Less than a minute (6 / 10%), 1 minute to less than 10 minutes (13 / 21%), 10 minutes to less than 1 hour (3 / 5%), 1 hour to less than 10 hours (9 / 14%), 10 hours to less than 1 day (8 / 13%), 1 day to less than 1 week (11 / 17%), I will wait long as it takes (13 / 21%)

Continued on next page

[29]One participant was not directed to this question after choosing solely "Other" in II.2. This applies also for II.7.

Table 3 – continued from previous page

SURVEY QUESTION	MAPPING TO RQ	n	RESPONSES (TOTAL / PERCENTAGE OF PARTICIPANTS)
III.5	RQ3	63	In a lab setting with simulated annotators (36 / 57%), In an applied setting with human annotators (48 / 76%)
IV.1∗∗[30]	RQ3	48	*free text answer* (47 / 98%), No answer (1 / 2%)
IV.2[31]	RQ3, RQ5	48	Classic (11 / 23%), Neural Network (5 / 8%), BERT era Language Models such as BERT, GPT-2, RoBERTa, ... (24 / 50%), Large Language Models such as Phi-3, Llama, Mistral, Gemma, Gemini, GPT-3/4, ... (2 / 4%), Do not know (2 / 4%), Other (4 / 8%)
		33	Classic (3 / 9%), Neural Network (4 / 12%), BERT era Language Models such as BERT, GPT-2, RoBERTa, ... (20 / 61%), Large Language Models such as Phi-3, Llama, Mistral, Gemma, Gemini, GPT-3/4, ... (2 / 6%), Do not know (2 / 6%), Other (2 / 6%)
IV.3	RQ3, RQ5	48	Uncertainty-based (20 / 42%), Disagreement-based (6 / 13%), Gradient information-based (2 / 4%), Performance prediction-based (5 / 10%), Density-based (1 / 2%), Discriminative-based (0 / 0%), Batch diversity (1 / 2%), Hybrid (7 / 15%), Do not know (4 / 8%), Other (2 / 4%)
		33	Uncertainty-based (13 / 39%), Disagreement-based (5 / 15%), Gradient information-based (2 / 6%), Performance prediction-based (4 / 12%), Density-based (1 / 3%), Discriminative-based (0 / 0%), Batch diversity (1 / 3%), Hybrid (3 / 9%), Do not know (3 / 9%), Other (1 / 3%)
IV.4∗∗	RQ3	48	We did not retrain (0 / 0%), Less than a minute (8 / 17%), 1 minute to less than 10 minutes (8 / 17%), 10 minutes to less than 1 hour (6 / 13%), 1 hour to less than 10 hours (8 / 17%), 10 hours to less than 1 day (6 / 13%), 1 day to less than 1 week (6 / 13%), Do not know (6 / 13%)
		33	We did not retrain (0 / 0%), Less than a minute (3 / 10%), 1 minute to less than 10 minutes (6 / 21%), 10 minutes to less than 1 hour (6 / 21%), 1 hour to less than 10 hours (4 / 14%), 10 hours to less than 1 day (5 / 17%), 1 day to less than 1 week (5 / 17%), Do not know (4 / 0%)
IV.5	RQ3, RQ5	48	Evaluation of the learned model on a held-out gold standard (15 / 31%), Money and/or time available for annotation were depleted (18 / 38%), All relevant documents annotated (4 / 8%), The stopping time was indicated by an algorithmic stopping criterion (6 / 13%), Do not know (2 / 4%), Other (3 / 6%)
		33	Evaluation of the learned model on a held-out gold standard (10 / 30%), Money and/or time available for annotation were depleted (13 / 39%), All relevant documents annotated (2 / 6%), The stopping time was indicated by an algorithmic stopping criterion (4 / 12%), Do not know (1 / 3%), Other (3 / 9%)
V.1∗[32]	RQ3, RQ5	48	Automatic speech recognition (0 / 0%), Coreference resolution (1 / 2%), Chunking (0 / 0%), Information extraction (5 / 10%), Language generation (0 / 0%), Language understanding ((0 / 0%)), Machine translation (3 / 6%), Morphological analysis (0 / 0%), Named entity recognition (5 / 10%), Part-of-speech tagging (0 / 0%), Question answering (0 / 0%), Relation extraction (1 / 2%), Semantic similarity (0 / 0%), Sentiment analysis (3 / 6%), Syntactic parsing (1 / 2%), Summarization (0 / 0%), Text categorization (22 / 46%), Word segmentation (0 / 0%), Word sense disambiguation (0 / 0%), Other (7 / 15%)
V.2∗[33]	RQ3, RQ5	48	Arabic (4 / 8%), English (32 / 67%), French (5 / 10%), German (10 / 21%), Hindi (1 / 1%), Mandarin (1 / 2%), Spanish (6 / 13%), Other (16 / 33%)
V.3∗	RQ3, RQ5	48	*Indication of hours* (22 / 46%), Do not know (26 / 54%)
V.4∗	RQ3, RQ5	48	*free text answer* (48 / 100%)
V.5∗	RQ3	48	Annotation was done by the project coordinators or other project members (20 / 42%), Domain experts (29 / 60%), Non-domain experts (12 / 25%), Other (3 6%)
V.6	RQ3	48	Yes (25 / 52%), No (23 / 48%)
		33	Yes (17 / 52%), No (16 / 48%)

Continued on next page

[30] The years of project start in the free text answer are: 2020–2024 (33), 2015–2019 (11), 2010–2014 (2), 2005–2009 (1).
[31] For the paper, we refined the wording by replacing "classic models" with "traditional models".
[32] We updated the 2009 task selection (Tomanek and Olsson, 2009) based on http://nlpexplorer.org/ and Hou et al. (2021).
[33] We adopted the language selection for this question from Tomanek and Olsson (2009).

Table 3 – continued from previous page

SURVEY QUESTION	MAPPING TO RQ	n	RESPONSES (TOTAL / PERCENTAGE OF PARTICIPANTS)
V.7[34]	RQ3	25	ActiveAnno (1 / 4%), Argilla (4 / 16%), AWS Sagemaker Ground Truth (0 / 0%), AWS Comprehend (0 / 0%), BioQRator (0 / 0%), INCEpTION (2 / 8%), Labelbox (0 / 0%), Label Studio (1 / 4%), MAT (0 / 0%), Potato (0 / 0%), Prodigy (2 / 8%), Other (15 / 31%)
		17	ActiveAnno (1 / 6%), Argilla (4 / 24%), AWS Sagemaker Ground Truth (0 / 0%), AWS Comprehend (0 / 0%), BioQRator (0 / 0%), INCEpTION (0 / 0%), Labelbox (0 / 0%), Label Studio (1 / 6%), MAT (0 / 0%), Potato (0 / 0%), Prodigy (2 / 12%), Other (9 / 53%)
V.8∗	RQ3	48	Deleted them (6 / 13%), Assigned one of two labels (14 / 29%), Not applicable, each instance had at most one annotation. (15 / 31%), Other (13 / 27%)
V.9∗	RQ3	48	No (27 / 56%), Yes (21 / 44%)
V.10∗	RQ3	21	Previously annotated examples were re-annotated. (12 / 57%), Previously annotated examples were left unchanged. (6 / 29%), Other (3 / 14%)
V.11	RQ3	48	Very successful (7 / 15%), Successful (35 / 73%), Neither successful nor unsuccessful (6 / 13%), Unsuccessful (0 / 0%), Very unsuccessful (0 / 0%)
		33	Very successful (6 / 18%), Successful (24 / 73%), Neither successful nor unsuccessful (3 / 9%), Unsuccessful (0 / 0%), Very unsuccessful (0 / 0%)
V.12	RQ3, RQ5	48	Very effective (12 / 25%), Effective (19 / 40%), Neither effective nor ineffective (14 / 29%), Ineffective (2 / 4%), Very ineffective (1 / 2%)
		33	Very effective (10 / 30%), Effective (12 / 36%), Neither effective nor ineffective (8 / 24%), Ineffective (2 / 6%), Very ineffective (1 / 3%)
V.13	RQ3, RQ5	36	In retrospect, I had insufficient expertise/knowledge. (6 / 17%), Overhead in setting up an Active Learning-based annotation environment. (10 / 28%), Lack of suitable annotation tools that easily integrate Active Learning. (11 / 31%), In retrospect, Active Learning did not meet my project's specific requirements. (2 / 6%), Active Learning created a sampling bias in the corpus. (6 / 17%), Active Learning did not reduce annotation cost. (4 / 11%), Active Learning did not work well in my scenario. (7 / 19%), My upfront estimation of the impact on reducing annotation costs was not accurate. (5 / 14%), The dependency of the model from the created dataset and vice versa. (7 / 19%), Other (8 / 22%)
		23	In retrospect, I had insufficient expertise/knowledge. (3 / 9%), Overhead in setting up an Active Learning-based annotation environment. (7 / 21%), Lack of suitable annotation tools that easily integrate Active Learning. (8 / 24%), In retrospect, Active Learning did not meet my project's specific requirements. (2 / 6%), Active Learning created a sampling bias in the corpus. (5 / 15%), Active Learning did not reduce annotation cost. (3 / 9%), Active Learning did not work well in my scenario. (7 / 21%), My upfront estimation of the impact on reducing annotation costs was not accurate. (3 / 9%), The dependency of the model from the created dataset and vice versa. (5 / 15%), Other (4 / 12%)
V.14	RQ3, RQ5	48	No (2 / 4%), Yes (42 / 88%), Other (4 / 8%)
		33	No (2 / 6%), Yes (28 / 85%), Other (3 / 9%)
V.15	RQ3, RQ5	2	Active Learning has become obsolete nowadays. (0 / 0%), Active Learning is a useful concept in theory but does not work well enough in practice. (2 / 100%), Other (0 / 0%)
		2	Active Learning has become obsolete nowadays. (0 / 0%), Active Learning is a useful concept in theory but does not work well enough in practice. (2 / 100%), Other (0 / 0%)
VI.1	RQ4	63	Strongly agree (22 / 35%), Agree (29 / 46%), Neither agree nor disagree (10 / 16%), Disagree (1 / 2%), Strongly disagree (1 / 2%)
VI.2	RQ4	63	Strongly agree (15 / 24%), Agree (27 / 43%), Neither agree nor disagree (17 / 27%), Disagree (2 / 3%), Strongly disagree (2 / 3%)
VI.3	RQ4	63	Strongly agree (15 / 24%), Agree (22 / 35%), Neither agree nor disagree (19 / 30%), Disagree (6 / 10%), Strongly disagree (1 / 2%)

Continued on next page

[34] We based the list of annotation tools on Borisova et al. (2024). For $n = 48$, in addition to 7 self-built solutions, participant-added options emphasize further relevant tools: LabelSleuth (4; Shnarch et al., 2022a), AL Toolbox (2; Tsvigun et al., 2022b), ALAMBIC (1; Nachtegael et al., 2023), and ALANNO (1; Jukić et al., 2023). For $n = 33$, in addition to 4 self-built solutions, participant-added options emphasize further relevant tools: LabelSleuth (3), ALAMBIC (1), and ALANNO (1).

Table 3 – continued from previous page

SURVEY QUESTION	MAPPING TO RQ	n	RESPONSES (TOTAL / PERCENTAGE OF PARTICIPANTS)
VI.4	RQ4	63	Strongly agree (11 / 17%), Agree (12 / 19%), Neither agree nor disagree (28 / 44%), Disagree (9 / 14%), Strongly disagree (3 / 5%)
VI.5	RQ4	63	*free text answer* (26 / 41%), No answer (37 / 59%)
VII.1	background	144	Academia (107 / 74%), Industry (45 / 31%), Governmental organization (7 / 5%), Other (2 / 1%)
VII.2	background	144	Linguistics (27 / 19%), Computational Linguistics (56 / 39%), Computer Science/Informatics (88 / 61%), Engineering (23 / 16%), Mathematics (10 / 7%), Social Sciences (14 / 10%), Other (6 / 4%)
VII.3	background	144	Economics (4 / 3%), Education (44 / 31%), Financial sector (5 / 3%), Health services sector (12 / 8%), Humanities (24 / 17%), Information and communications technology (79 / 55%), Legal sector (8 / 6%), Public service sector (11 / 8%), Society and politics (7 / 5%), Other (18 / 13%)
VII.4	background	144	Up to 1 year (1 / 1%), 1-2 years (11 / 8%), 3-5 years (47 / 33%), 5-10 years (55 / 38%), More than 10 years (30 / 21%)
VII.5	background	144	*Indication of country* (143 / 99%), Other (0 / 0%), No answer (1 / 1%)
VII.6∗	background	144	*free-text answer* (22 / 15%), No answer (122 / 85%)
VII.7∗∗	background	144	Social media (16 / 11%), Personalized email invitation (33 / 23%), Mailing list (73 / 51%), Friend, colleague, or peer shared it (13 / 9%), Other (6 / 4%), No answer (3 / 2%)

Table 3: Overview of response selection. Each survey question is linked to a research question. We indicate the number of participants n per question and their distribution across predefined response options. In case of questions described in Section 3.4, we provide both the full set of respondents ($n = 48$) and the set of respondents that started the reported AL project in 2020 or later ($n = 33$, marked in bold). Free-text responses are omitted for scope but can be seen in the dataset, provided in the supplementary material. $*$: marks questions not discussed in the main body of the paper, $**$: marks questions only discussed partly.