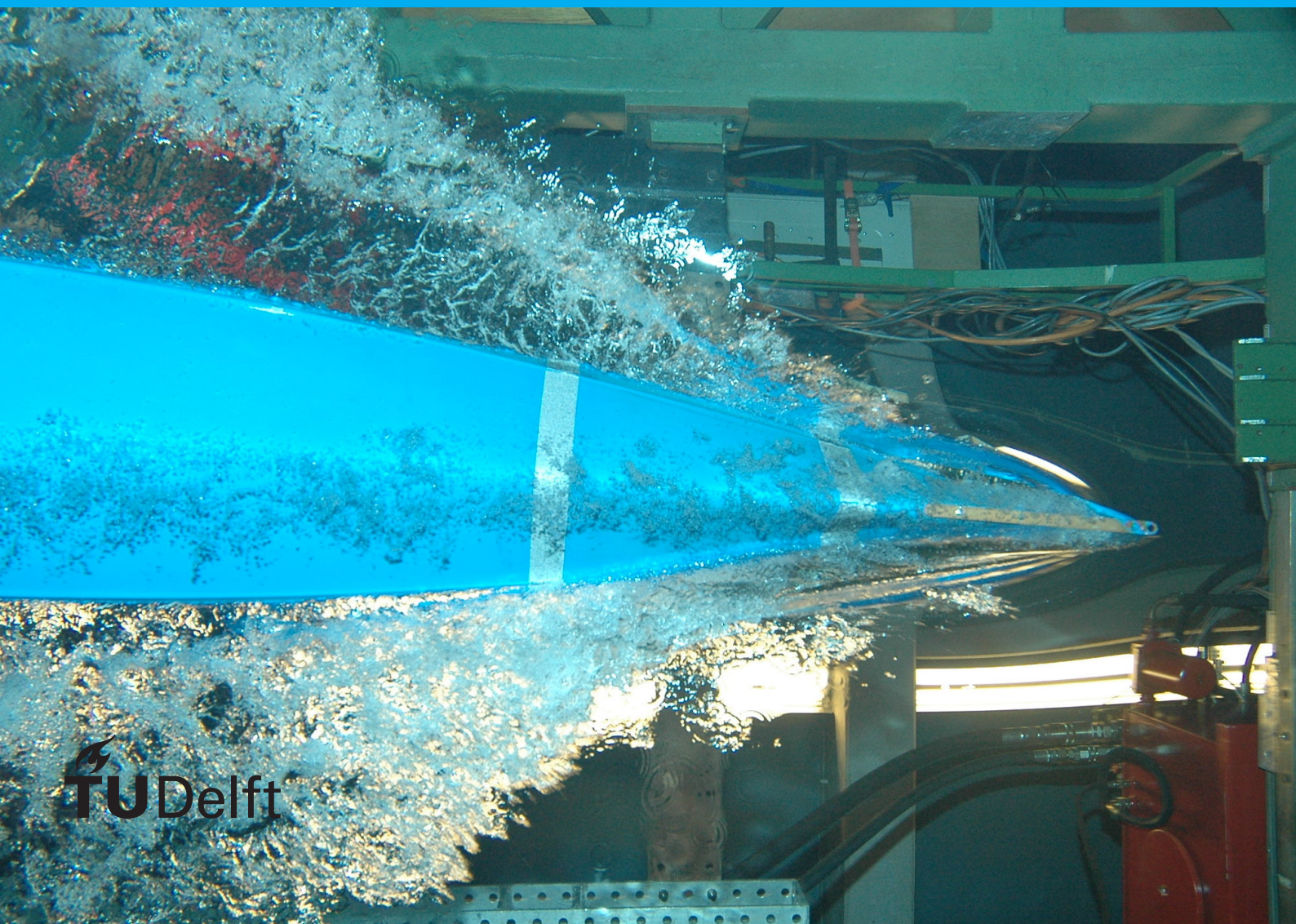
Human-AI Collaboration for Policy Document Annotation

Case of the CoronaNet Dataset

Ye Yuan

Technische Universiteit Delft



Human-AI Collaboration for Policy Document Annotation

Case of the CoronaNet Dataset

by

Ye Yuan

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 24, 2022 at 15:00 PM.

Student number:	5218683	
Project duration:	September 1, 2021 – August 3, 2022	
Thesis committee:	Prof. dr. ir. Geet-Jan Houben,	TU Delft
	Dr. ir. Ujwal Gadiraju,	TU Delft, supervisor
	Dr. ir. Pradeep Murukanniah,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

It has been a wonderful experience to complete my master's project thesis in the past nine months. I thank everyone who helped and accompanied me during the dissertation process. First of all, I would like to thank my daily supervisor Dr. Ir. Ujwal Gadiraju guided me into this exciting topic and encouraged me throughout the thesis process. I would also like to thank Prof. Geert-Jan, who gave me valuable feedback on critical points. Second, I would like to thank all my family and friends who accompanied and encouraged me until the end of this dissertation journey.

Ye Yuan
Delft, August 2022

Contents

Abstract

Since the Covid-19 crisis, countries around the world have responded with various policies, mostly with little success. It is a consensus that the new coronavirus cannot be eradicated. Therefore, in order to help governments make trade-offs and control the epidemic without affecting the economy, many research institutions or non-profit organizations have been annotating the policy documents of governments around the world and building relevant data sets. These data provide practical reference for future decision-making in countries around the world and open the possibility for wider application. Policy annotation requires a lot of heavy and repetitive work, which is not only time-consuming but also inefficient. Using NLP models to assist the annotation work was not been explored. However, due to the limitations of artificial intelligence development and the diversity of annotating tasks, it is not feasible to completely replace manual annotating work with artificial intelligence models. Therefore, the efficiency of annotating tasks can be significantly improved through the Human-AI Collaborative Workflow approach, combining the respective advantages of humans and AI. However, it is likely to have adverse effects due to various human or uncontrollable factors. For example, AI models cannot guarantee that every policy document can generate high-quality recommended answers, the UI of workflow systems can mislead annotators, etc. On the one hand, the quality of the annotation results generated by the AI model directly determines the accuracy of the annotation task. On the other hand, presenting the halfway product generated by AI to the annotator is also an essential factor in deciding the annotation task's efficiency. But can the idea of "Human-AI collaboration" improve labeling efficiency and reduce costs? We investigated whether this approach was positive, negative, or irrelevant to annotator productivity. The control group annotated manually in the experiment, while the experimental group annotated with the help of a Human-AI collaborative workflow approach. Our research found that annotators tend to trust AI results. This results in a positive correlation between the quality of AI generated semi-finished products and the accuracy of the labeling task. That is, high-quality AI recommendations can improve the accuracy of annotation results and shorten the completion time. However, low-quality AI recommendations reduce the accuracy of the annotation results and take longer to complete than pure human annotation because the annotator needs extra time to trade off the AI suggestions.

1

INTRODUCTION

Documentation annotation is indispensable for many industries, such as medical, food industries, etc. Accurate documentation annotations are critical to the long-term operation and growth of an industry. Since the COVID-19 crisis, countries around the world urgently need a high-quality dataset of policy documents to guide government behavior to control epidemics and maintain economic stability. But the sheer volume of policy documents spanning different languages and countries requires expertise. Maintaining a document annotation effort requires a large organization or institution and involves policy documentation on a global scale.

There are many datasets for the response to Covid-19; for example, the European Centre for Disease Control and Prevention maintains many datasets on vaccination data, national responses, hospital and ICU admission rates, etc. One of them is CoronaNet, a policy issued in response to governments worldwide during the Covid-19 pandemic. In this article, we study CoronNet as an example.

Previous research has analyzed the impact of highlighting on the productivity of annotators. Wu et al. have shown that text highlighting can decrease reading time, but inappropriate or irrelevant text highlighting can have the opposite effect [39]. Alagarai Sampath presents an approach to assist crowd workers in digitization works [2]. Schaekermann et al. requested highlights as evidence to support judgments [35]. Jorge Ramirez et al. studied the impact of ML highlighting different qualities to the text annotators in a true or false annotation task [34]. However, there are a lot of annotation problems like generating summaries, filling in the blanks, choosing, etc. There are no studies on the effect of highlighted answers on annotator productivity in these tasks. There is a lot of research on Human-AI collaboration and Human-AI interaction. However, there is currently no research on combining multiple AI models to form a Human-AI Collaborative Workflow System for multiple mixed types of annotating tasks. This thesis will fill this gap and propose a general Human-AI collaborative workflow for document annotation based on the idea of Human-AI collaboration and the human-in-the-loop system.

The research in this paper focuses on whether and how AI models can improve the productivity of an annotator. The AI models involve text summarization, question answering, and document embedding. We mainly study the following metrics: accuracy and completion time. More specifically, the completion time directly determines the cost of time-based annotation tasks. The following two questions are the research questions of this paper:

1. **Does the accuracy increase, decrease or stay the same with the support of the Human-AI collaborative workflow?**

2. **Do completion times increase, decrease, or stay the same, supported by Human-AI collaborative workflows?**

In this paper, we hired 192 annotators and divided them into three groups. One group annotated a policy document purely by humans, one group annotated the same document with the support of high-quality AI recommendations, and the last group with the help of low-quality AI recommendations. And everyone needs

to answer a questionnaire about the user experience of the Human-AI collaborative workflow system at the end of the task.

Experimental results show that the efficiency of the annotator is positively correlated with the quality of the recommendations provided by the AI. It means that high-quality AI recommendations can improve the accuracy of annotators' results and reduce working hours, and vice versa. In addition, the statistical results of the questionnaire showed that among the three experimental groups, the group that was provided with high-quality AI recommendations believed that the system could provide users with higher performance and less mental, physical, and time demands. Additionally, annotators in this group felt that using the system to do annotate tasks required the least effort and experienced minor frustration compared to the other two groups.

This main contributions of this thesis are:

1. This thesis reveals that a workflow model that combines AI recommendations and multiple ML models is practical and effective for improving crowdsourced the annotation tasks.
2. This thesis contributed a Human-AI collaborative workflow based application for crowdsourcing annotation tasks.

2

BACKGROUND AND RELATED WORK

In recent years, there has been a lot of work on Human-AI collaboration workflows. In this section, references will be made to those works that are most relevant to the subject. We will focus on this thesis's three most relevant areas: Human-AI Collaboration, Human-In-The-Loop Workflows, and Document Annotation.

2.1. Human-AI Collaboration

Pure human-based systems and pure AI-based systems are already widely used in every aspect of our lives, each with different shortcomings and limitations when working alone. If humans and AI work together, they can empower each other to achieve better results. The main challenge for a standalone AI agent is how to achieve its goals effectively and flawlessly. However, in teams where humans and AI help each other achieve team goals, the challenge is not limited to the goals themselves, but should also be able to reason about human behavior. There are various jobs in various fields trying to get humans and AI to work together as a team for better results. While the works in this field focus on different challenges, there is a lack of coherence between them, and it is difficult to see a clear connection between these works. James A. Crowder et al. describe in detail the architecture and algorithms of an Intelligent Information Software Agent (ISA) cognitive system that facilitates collaborative communication between humans and AI systems [10]. Juan Liu et al. proposed a data analysis system of Human-AI collaboration, in which user workflows are recorded and common workflow patterns are learned using graph analysis, information scent, and example-based learning techniques. The system automates tedious processing steps to increase analyst productivity, which can provide recommendations based on expert user workflows [32]. Gadiraju et al. discover new ways of managing task assignment and delivery, coordinating multiple populations in collaborative and competitive task execution, and new data analysis methods that can lead by exploring the best trade-offs between labs and populations, and through populations and a rich dataset resulting from mixed methods experiments [20]. (Hugo Scurto et al. 2018) This paper proposed a prototypical computational framework for music appropriation. The pedagogical potential of this framework is demonstrated in two of music applications they implemented [36]. (Yi-Ching Huang et al. 2019) A conceptual framework called "co-learning" is proposed. Users can learn and grow with AI partners in this framework over time, suggesting it can improve productivity and creativity in creative problem areas [26]. Gadiraju et al. identify opportunities in crowd computing to advance better AI techniques. They believe these advances require solving fundamental problems from a computational and interactive perspective, articulating a world where humans and AI can be seamless and mutually beneficial [18]. (Lanthao Benedikt et al. 2020) A human-in-the-loop artificial intelligence application is proposed to help government agencies generate statistics automatically. This enables humans to focus on value-added tasks that require flexibility and intelligence [3]. António Correia et al. presented a model that incorporates the core principles of human-machine symbiosis (HMS) into scientometric workflows. It is an initial design of a Human-AI enabled pipeline for performing scientometric analyses, leveraging the intersection between human behavior and machine intelligence [8]. Zehao Dong et al. proposed a deep graph neural network IDSP to incorporate gene-gene and gene-drug regulatory relationships into synergistic drug combination prediction [13]. Li Fei. proposed a systematic approach to design a repeatable PHM sys-tem based on Human-AI collaboration, which not only provides competitive performance, but also provides consistent and repeatable results under different operating conditions [30]. (Andy Coenen et al. 2021) Wordcraft is an AI-assisted editor

for story writing. It allows a writer and a dialog system collaborate to write a story [7]. Suzanne, Gadiraju et al. built an intelligent house recommendation system and conducted a 3-session, longitudinal study of 201 participants over a week. They found evidence suggesting that trust development is a slow process that evolves over multiple sessions and that first impressions of the intelligent system are highly influential [38].

2.2. Human-In-The-Loop Workflows

Most AI cannot learn autonomously at the current stage, and 90% of machine learning applications rely on supervised learning. Smart devices have learned more from human examples and feedback than from hard-coded rules in the past. These human-coded examples (training data) are used to train machine learning models and make those models more accurate at performing a given task. But programmers still need to create software systems that allow feedback from non-technical people, which raises one of the most critical questions in technology today: What is the right way for humans and machine learning algorithms to interact to solve problems? Annotation and active learning are the cornerstones of human-in-the-loop machine learning. Deep learning has achieved great success in various applications such as natural language processing, speech recognition, medical applications, computer vision, and intelligent transportation systems [12]. The great success of deep learning is attributed to larger models [5]. The scale of these models already contains hundreds of millions of parameters. These hundreds of millions of parameters allow the model to have more degrees of freedom, enough to have excellent description capabilities. Integrating prior knowledge into the learning framework effectively deals with sparse data since the learner does not need to generalize knowledge from the data [11]. More and more researchers have recently incorporated pretrained knowledge into their learning frameworks [6], [31], [25]. Sharifi et al. introduced a human-in-the-loop semantic analysis framework for large-scale characterization of unknown unknowns. It provides a rich, descriptive report of unknown unknowns and allows for more efficient and cost-effective detection than existing techniques[37]. Xin et al. propose a "human-in-the-loop" machine learning system that enables rapid iteration, response to feedback, introspection and debugging, and background execution and automation. It has made typical iterative workflows 10 times faster than competing systems [40]. Li et al. present a workflow perspective on AutoML. The first obtained workflow is fed to an execution engine, which executes the actions specified by the workflow, producing a set of ML results (which can be categorical labels or a table of predicted values). Finally, the machine learning results are returned to the user for debugging and analysis [28]. Bode et al. evaluated existing HITL AI systems for clinical use. The study's results showed that HITL and automatic DIA were significantly more accurate, with more minor deviations in standard errors [4]. Krano et al. proposed a production quality system that hundreds of people use to schedule thousands of actual meetings in a year. This research introduces a novel architecture that seamlessly combines automation, microtask, and macrotask execution to generate a responsive and scalable scheduling assistant that demonstrates its value through large-scale field deployment [9]. Alexander et al. conducted a study. The results show strong responder preferences against the algorithm, as most responders opt for a human opponent and demand higher compensation to reach a contract with autonomous agents [15].

2.3. Document Annotation

During the COVID-19 pandemic, governments and research institutions in various countries tried to establish a policy database related to COVID-19 treatment. This requires a lot of materials, money, and time for crowdsourcing staff to collect data and annotate them. The task is mainly about the text annotation. Therefore, a Human-AI workflow application based on document annotations was developed to help improve the efficiency of crowdsourcing tasks. It can respond to the Covid-19 pandemic and help improve the efficiency and accuracy of similar tasks in other fields. The quality of crowdsourcing document annotation tasks can be affected by a series of factors, including task clarity, work environments, cognitive biases, task complexity, task design and ordering, participant moods, user interface factors, participant behavior, and annotation strategies. Gadiraju et al. show that task clarity is coherently perceived by crowd workers, and is affected by the type of the task [21]. Gadiraju et al. find crowd workers are embedded in a variety of work environments which influence the quality of work produced. The crowdsourcing with the best work environments is ModOp, which results in reducing the cognitive load of workers, thereby improving their user experience without effecting the accuracy or task completion time [19]. Christoph Hube et al.'s study results reveal that workers with strong opinions tend to produce biased annotations. Such bias can be mitigated to improve the overall quality of the data collected. Experienced crowd workers also fail to distance themselves from their own opinions to provide unbiased annotations [27]. Gadiraju et al. present the reader with an overall under-

standing of the kinds of experiments that can benefit from the virtues of crowdsourcing and the cases that are less suitable for the same [20]. Gadiraju et al. proposed a method for worker pre-selection leads to a higher quality of results when compared to the standard practice of using qualification or pre-screening tests. For image transcription tasks this method resulted in an accuracy increase of nearly 7% over the baseline and of almost 10% in information finding tasks, without a significant difference in task completion time [22]. Alan Aipe et al. reveal the impact of chaining tasks according to their similarity on worker accuracy and their task completion time [1]. Sihang Qiu et al.'s experimental results show that conversational interfaces can be effective in engaging workers, and a suitable conversational style has potential to improve worker engagement [33]. A study by Lei Han et al. shows that the presence of frequently used shortcut patterns can speed up task completion, thus increasing the hourly wage of efficient workers [24]. Wu et al. pointed out that text highlighting can reduce people's reading time [39]. Gier et al. pointed out that low-quality text highlighting will have a negative impact on human reading [23]. Alagarai Sampath et al. shows that highlighting can help workers in digitization tasks [2]. Schaekermann et al. use highlighting as evidence to support judgment[35]. Jorge et al. verified that in text classification tasks, ML Highlighting can help crowd workers improve efficiency, but it cannot increase the accuracy of crowdsourcing tasks [34]. Based on this research, this thesis will contribute to it by building a Human-AI Collaborative Workflow that supports more task types and more rich AI suggestions.

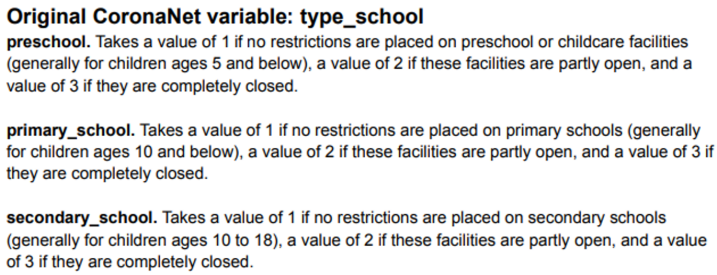
3

THE CORONANET DATASET

Although computational text analysis promises less error-prone, expensive, and resource-intensive investigations of policy text, the extent to which this promise can be fulfilled remains unclear. This project aims to address this gap by evaluating the accuracy and precision of various natural language processing and text analytics pipelines and techniques in classifying and clustering policy text. The project is based on the CoronaNet dataset, a publicly available source of over 50,000 policy announcements made by over 190 countries to respond to the COVID-19 pandemic. This dataset has been collated and coded by over 500 researchers. The project will involve developing various pipelines for classifying and clustering the policy announcements recorded in this dataset and comparing the results obtained with hand coding of policy announcements. The findings of this project will advance research on the application of machine learning to public policy text and contribute to the creation of best practices for automated analysis of public policy documents.

Currently, CronaNet is maintained and annotated by a group of people. The annotation method is now purely manual. These annotators (Data Analysts) need to read the policy document carefully, extract meaningful information, and then analyze the information. Finally, according to the requirements of 'Index Codebook' [See Figure 3.1], assign values to the various fields of CronaNet. This process is cumbersome and time-consuming, and the final result needs to be verified by specialized personnel (Data Validators). If an incorrect place is found, it must be returned to annotators to modify it again. This process requires multiple iterations.

This situation increases the cost and severely limits the efficiency of crowd workers. Therefore, this article first provides a workflow-based idea. Based on this idea, an AutoAI application combining ML highlighting, text summarization, Multiple-Choice, and QA is developed to help reduce cost and improve the efficiency of crowd workers.



Original CoronaNet variable: type_school

preschool. Takes a value of 1 if no restrictions are placed on preschool or childcare facilities (generally for children ages 5 and below), a value of 2 if these facilities are partly open, and a value of 3 if they are completely closed.

primary_school. Takes a value of 1 if no restrictions are placed on primary schools (generally for children ages 10 and below), a value of 2 if these facilities are partly open, and a value of 3 if they are completely closed.

secondary_school. Takes a value of 1 if no restrictions are placed on secondary schools (generally for children ages 10 to 18), a value of 2 if these facilities are partly open, and a value of 3 if they are completely closed.

Figure 3.1: An example in the 'Index Codebook'.

4

A SYSTEM FOR HUMAN-AI COLLABORATIVE ANNOTATION

This section illustrates the implementation of the idea of the Human-AI collaborative workflow for the policy document annotation.' The user interface of the system is presented in Figure 4.2. This system consists of two main parts, one for pure human annotation and one for Human-AI collaboration. The two parts share the exact structure of the user interface, the policy text display on the right and the question answering box on the left. The details are described in the following:

4.1. From policy document to "description" in CoronaNet

From the CoronaNet Codebook found on the official website, you can see that the first step is Data Collection. The CoronaNet Research Project uses two Machine Learning companies, Jataware and Overton. The main data collected by them are news articles related to COVID-19 around the world. After collecting the data, Jataware processes the data in two steps. The first processing step is to judge whether each article collected is related to COVID-19. The second processing step is to generate a list of options for each field, for example, the list of options generated for the policy type field are as follows (see Figure 4.1):

Responses:

- Declaration of Emergency: The head of government declares a state of national emergency.
- Lockdown: Targets of the policy are obliged shelter in place irrespective of potential likelihood of COVID-19 transmission and are only allowed to leave their shelter for specific reasons.
- Curfew: Government policies that limit domestic freedom of movement to certain times of the day.
- Quarantine: Targets of the policy are obliged to isolate themselves for at least 14 days because there is reason to suspect a person is infected with COVID-19.
- External Border Restrictions: Government policies which reduce the ability to access ports of entry or exit to or from different governmental jurisdictions.
- Internal Border Restrictions: Government policies which reduce the ability to move freely within the borders of the initiating government.
- Restrictions of Mass Gatherings: Government policies that limit the number of people allowed to congregate in a place. Please enter the number in the text entry.
- Closure and Regulations of Schools: Government policy which regulates educational establishments in a country. This may include: closing an educational institution completely, allowing an educational institution to open with certain conditions; allowing an educational institution to stay open without conditions.

Figure 4.1: The list of options generated for the policy type field.

These lists will be integrated into a Qualtrics survey with survey questions and distributed to annotators. Then annotators answer these questions based on the policy content.

According to the above description, annotators need to read the original policy when annotating. However, the original text provided by Jataware is not concise, and annotators still need to read the entire document.

If the important sentences can be marked in the original text by Text Extraction model, annotators only need to read these summaries. (According to a study in terms of Text Summarization [14], the length of a summary is generally only one third of its original text.) it will improve the efficiency of annotators.

4.2. Annotating the blank filling questions

For open questions, for example, “country” in CoronaNet: "From what country does this policy originate from?". It is not provided with a response list, and annotators need to infer the answer based on the context of the policy. Thus, a QA model is applied here.

4.3. Annotating the multiple-choice questions

Some questions only require RAs to select an option from the response list provided by Code Book (for example, the Policy Types mentioned above). Such annotation tasks are done by a QA model first, and then pick the option which is has the highest cosine similarity score with the answer generated by QA model

4.4. Human annotation

1. Task 1, filling in the summary of the policy.

   In this annotation task, an annotator needs to summarize a policy (cf.A). The annotator should first read the policy on the right side and fill the summary in the text box (cf.B). Click the ‘Save’ button and ‘Go to annotation’ to the next task (cf.C).

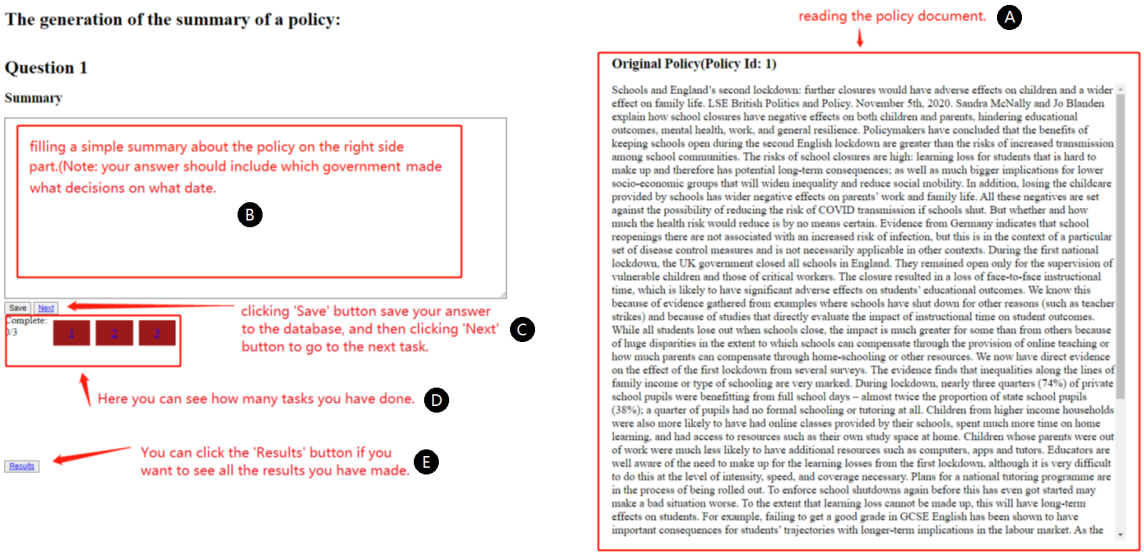


Figure 4.2: The annotating window for generating a summary.

2. Task 2, selecting an option based on the question and clarification of the property of the data-set (cf.B).

   In this task, an annotator needs to first read the policy text on the right side and the property's question and its clarification (cf.A), then select an option (cf.C) by clicking a radio button and save the answer (cf.D).

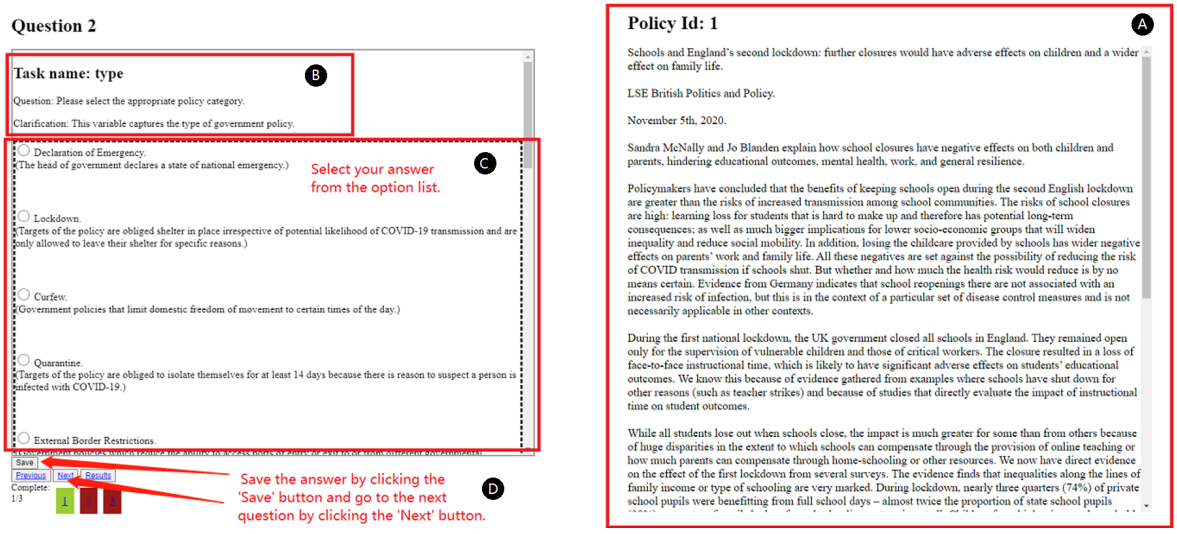


Figure 4.3: The annotating window for the multi-choice task.

3. Task 3, filling in the answer based on the question and clarification of the property of the data-set (cf.B).

   In this task, an annotator needs to first read the policy text on the right side (cf.A) and the property's question and its clarification (cf.B), then fill the answer in the text box (cf.C) and save the answer. Click the 'Save' button (cf.D).

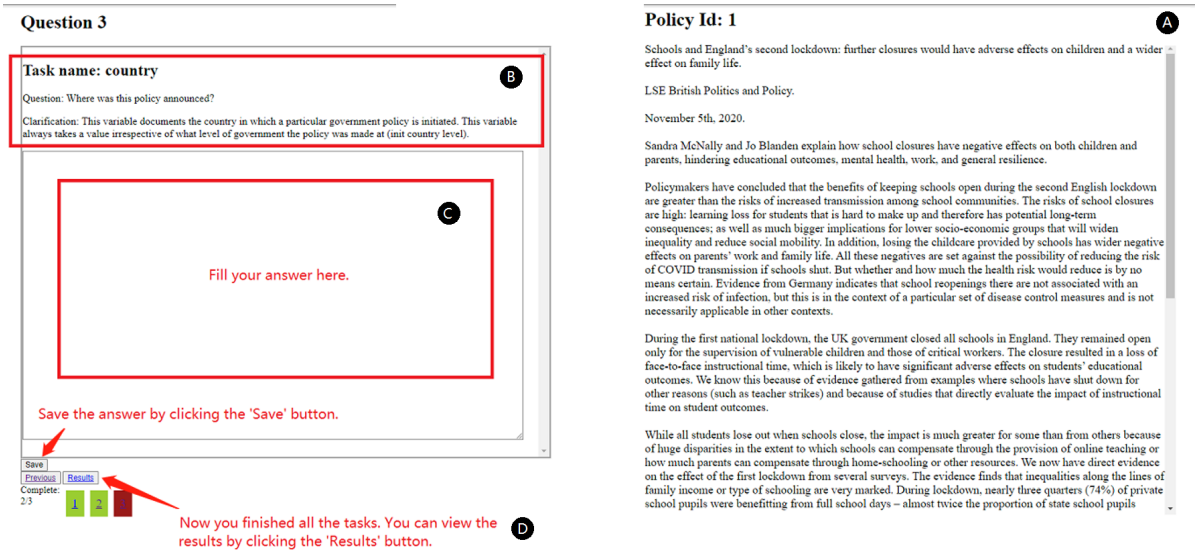


Figure 4.4: The annotating window for the fill-in task.

4.5. Human-AI collaborative annotation

In this part, the user of the system are provided with supports from the AI models.

1. Task 1, filling in the summary of the policy. The system displays the highlighting for summary on the right window.

In this annotation task, an annotator needs to summarize a policy. The annotator should first read the policy on the right side (cf.A) and check if the highlighted part is the wanted answer and do some modifications (cf.B). Finally, click the 'Save' button to save the answer to the database (cf.C).

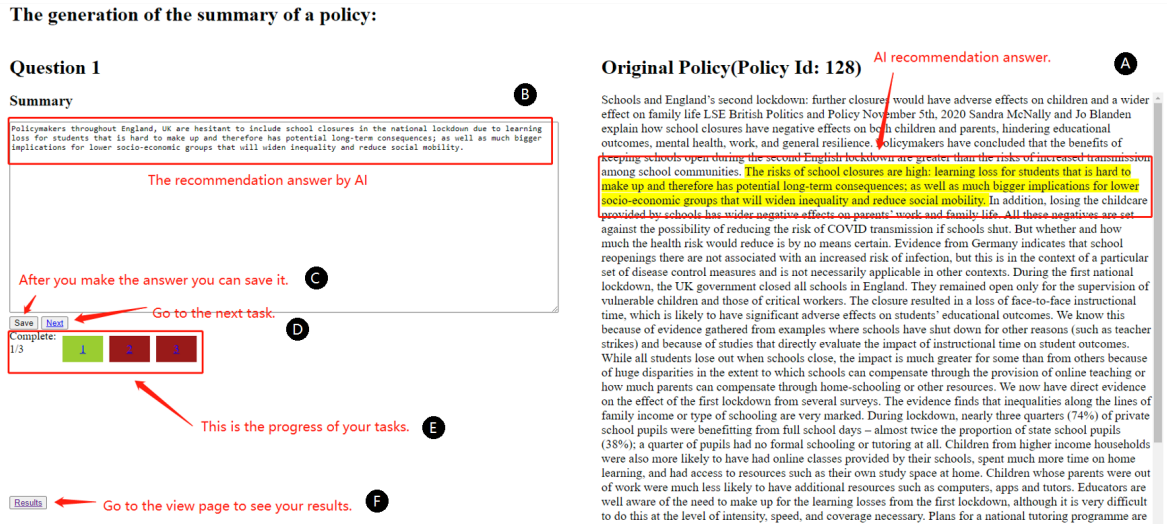


Figure 4.5: The annotating window for generating a summary.

2. Task 2, selecting an option based on the question and clarification of the property of the data-set. The system displays the highlighting for the AI-recommended option on the right window, and the confidence score under the AI-recommended option.

   In this task, an annotator need to do a selection. The annotator needs to read the policy text (cf.A). And then, read the question and the clarification of the property on the left side (cf.B). Finally, based on the highlight and AI recommendation option, select an answer from the options (cf.C) and click the 'Save' button (cf.E). Click the 'Next' button to go to the next task (cf.E). (Note: the AI Recommendation label can be red or green, which means that it is low and high confidence respectively (cf.D). So, the annotator should pay more attention when it is red.)

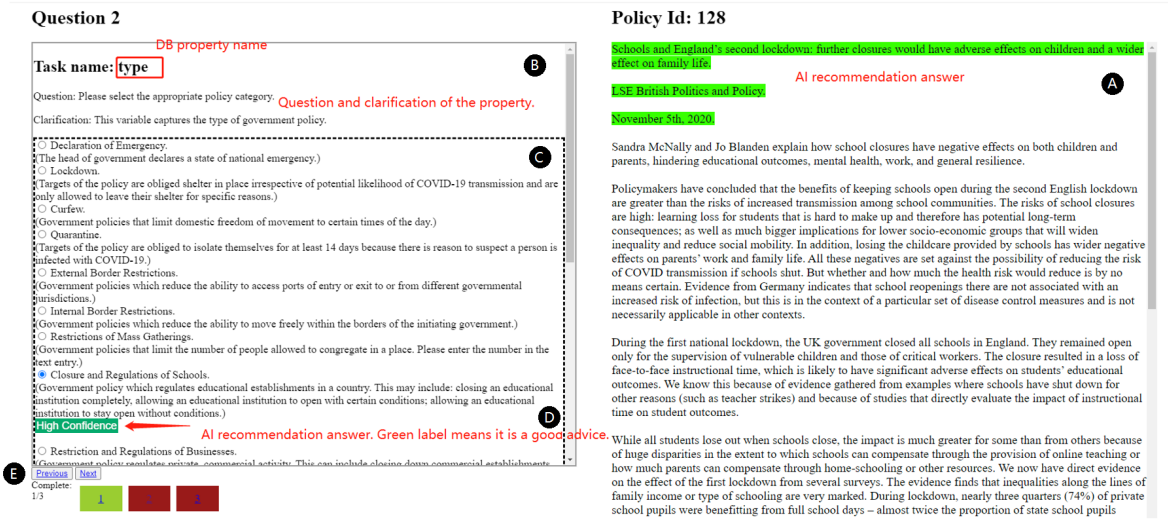


Figure 4.6: The annotating window for the multi-choice task.

3. Task 3, filling in the answer based on the question and clarification of the property of the data-set. The system provides five candidates and the corresponding highlighting for the user. When a user select a

candidate, it will be fill in the blank and the highlighting is displayed on the right window.

In this task, an annotator needs to read the policy text (cf.A) and the questions and the clarification (cf.B) of the property. And based on one of the candidate answers (cf.C), the annotator fill the answer in the text box (cf.D). Click the 'Save' button (cf.F). Now, all four tasks are done. [Note: when the the radio button of the AI recommendations is clicked, the system highlights the corresponding answer on the right side and fill the recommendation in the text box.]

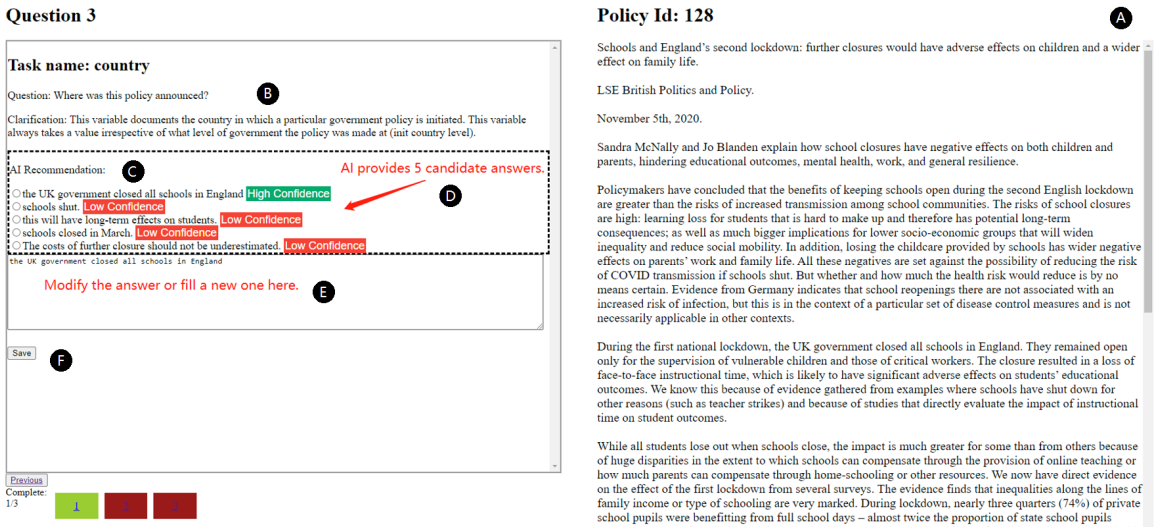


Figure 4.7: The annotating window for the fill-in task.

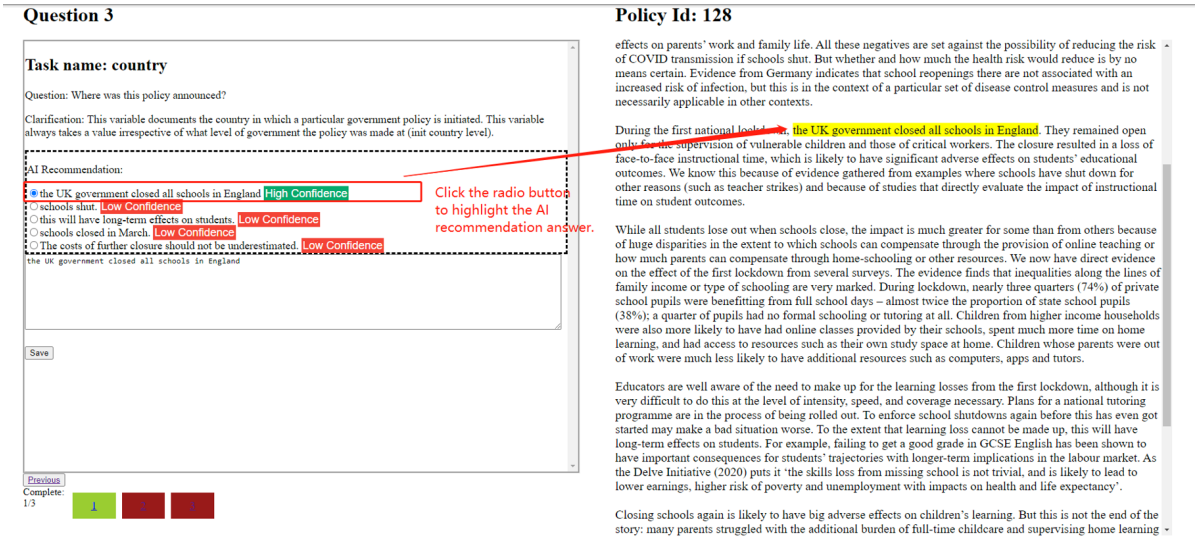


Figure 4.8: The annotating window for the fill-in task.

4.6. Workflow

This section is a detailed introduction to the Human-AI Collaboration Workflow. Figure 4.9 is the workflow chart of the Human-AI collaborative system. As can be seen from the figure, first, the annotator only needs to select a policy and one attribute of the dataset. Then, the system will extract the codebook rule and policy text corresponding to the attribute. Depending on the type of attribute, the system selects an appropriate AI model to generate answers and confidence levels. After this, the system generates the UI and adds the answer's highlighting and confidence scores, as shown in Figure 4.10. Finally, the system fills in the answer in the blanks of the UI. Confidence scores are suggestions given by the system by which annotators can check

the AI's answers. If the annotator does not find any errors, save the answer to the dataset. Otherwise modify the answer.

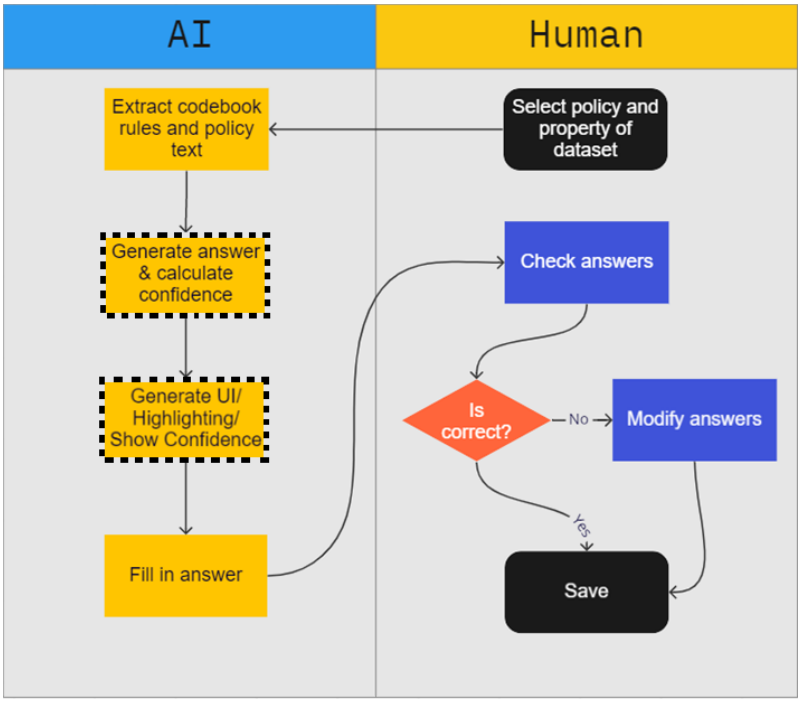


Figure 4.9: Workflow chart.

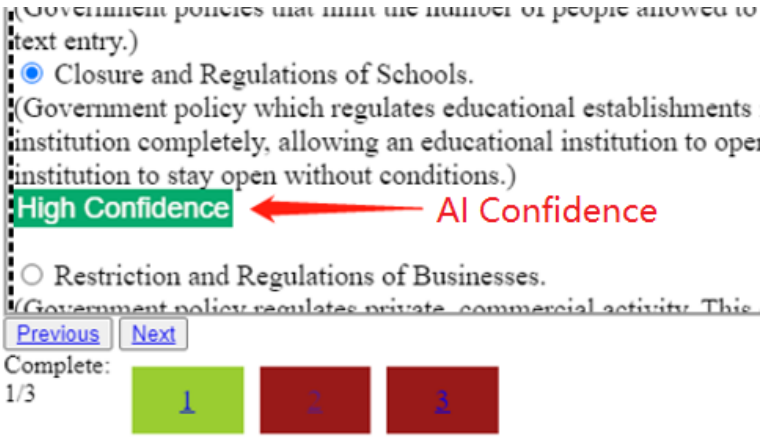


Figure 4.10: The UI of the AI confidence.

date_start	date_end	country	ISO_A3	ISO_A2	init_country_level	domestic_policy	province
2020-03-06		Afghanistan	AFG	AF	National	1	
2020-03-16		Afghanistan	AFG	AF	National	1	
2020-05-28		Afghanistan	AFG	AF	National	1	
2020-06-28		Afghanistan	AFG	AF	National	1	
2020-03-09		Afghanistan	AFG	AF	Municipal	1	Herat
2020-03-09		Afghanistan	AFG	AF	Municipal	1	Herat
2020-03-09		Afghanistan	AFG	AF	Municipal	1	Herat
2020-03-11		Afghanistan	AFG	AF	Provincial	1	Samangan
2020-03-11		Afghanistan	AFG	AF	Provincial	1	Samangan

Figure 4.11: An example of CoronaNet DB.

4.7. MODELS

4.7.1. Text summarization

The research of [34] summarizes the popular Text Summarization technologies up to 2021, including ML based Summarization Approaches, Neural Network based Summarization approaches, DNN based Approach (BART) and KG Augmented Summarization (SOTA). The performance (Rouge score) of the models involved is listed in Table 1. According to this list, BART (best performance) is chosen as the Text Summarization model of the system. The system applies BART, a pretraining sequence-to-sequence model, as a denoising auto-encoder. BART, consisting of BERT as an encoder and GPT as a decoder, performs well on different tasks including abstractive summarization and question answering. [29]

Model	**Rouge-1**	**Rouge-2**	**Rouge-L**
LEAD-3	39.20	15.70	35.50
SummaRuNNer	39.60	16.20	35.30
Abstractive Model	35.46	13.3	32.65
s2s 150k vocab	30.49	11.17	28.08
s2s 50k vocab	31.33	11.81	28.83
Pointer Generator	36.44	15.66	33.42
PTGEN + COV	39.53	17.28	36.38
BART	44.16	21.28	40.90
SOTA	43.93	20.37	40.48

Table 4.1: The Rouge score for different models.

4.7.2. Blank filling

For the blank filling questions, this Human-AI collective application uses BART as the QA algorithm and uses the description of this field in the Code Book as the input question. And combined with KG, generate an answer, this answer may be a sentence or a paragraph.

4.7.3. Multiple-Choice

For the multiple-choice questions problems, the model used is the same as QA. In addition, after the process mentioned in the QA section, the generated answer and each choice of the multiple-choice question are applied to calculate the co-sine similarity.

5

METHODOLOGY

In this section, we evaluate the Human-AI collaboration workflow. The purpose of the system is to annotate any document in various ways, including generating document summaries, filling in the blanks, and selecting the correct option based on the question. The system can generate halfway products of any type of document. To assess the approach, the system generates three types of user interfaces for an annotator: without AI's advice, bad advice, and good advice. We analyzed 20 policies with the system and picked one with good quality AI advice (https://blogs.lse.ac.uk/politicsandpolicy/schools-second-lockdown/) and one with bad advice (https://www.gov.scot/news/back-to-school-1/).

5.1. Data source

The data involved in this experiment are primary data collected from Prolific. The Human-AI Collaborative Workflow system of this paper was first deployed on SURF Cloud and then published on "Prolific". The specific release details will be detailed in subsequent chapters.

5.2. System Implementation

The implementation of Human-AI Collaborative Workflow System is based on the following frameworks and components: Flask 2.1.2, Flask-SQLAlchemy 2.5.1, Jinjia2 3.1.2, gensim 4.2.0, nltk 3.7, numpy 1.22.4, scikit-learn 1.1. 1. Sentence converter 2.2.0, torch 1.11.0 and sqlite3.

5.3. Experimental Setup

1. **Participants**

   'Prolific' is a well-known crowdsourcing site that provides powerful and flexible tools for online research. Tasks posted on Prolific are paid. Our assignments pay £8.04 an hour. The approved rate of task submissions is about 33.86% (192 out of 567).The geographic location of the participants was set to global to avoid regional bias. To enable crowdsourcing staff to provide high-quality answers, we are offering an additional £1 bonus to staff who get the job done accurately. Crowd workers performed annotation tasks according to the instructions provided (see Appendix A).

   A priori power analysis was performed using G*Power version 3.1.9.7 (Faul et al., 2007) to determine the minimum sample size required to test our research hypotheses. The results show that the sample size needed to detect a moderate effect is 80%, with a significance criterion of = .05, N = 192 for the T-Test. Therefore, a sample size of N = 192 was sufficient to test the research hypothesis. The detailed parameters of G-Power([16],[17]) are as follows:

T–Tests,
two independent means(two groups),
two tails,
effect size d=0.5,
$\alpha - errprob = 0.05$,
power=0.8,
allocation ratio N2/N1=1.

2. **Measures**

At the end of the annotation task, annotators were asked to answer a questionnaire. In this experiment, the questionnaire was built by Qualtrics, which is a popular online platform for research. The aim of this questionnaire is to investigate the user experience of the Human-AI Collaborative Workflow system. The questions involved in the questionnaire are based on the NASA TLX. NASA TLX is a famous multi-item questionnaire invented by Sandra Hart in 1980. TLX represents 'task load index', used for measuring the workload of perceived workload. It consists of six main metrics: mental demand, physical demand, temporal demand, performance, effort, and frustration. With these six metrics and corresponding questions, we investigate the user experience of the system. The full questionnaire is shown in Appendix B. The questions involved are as follows:

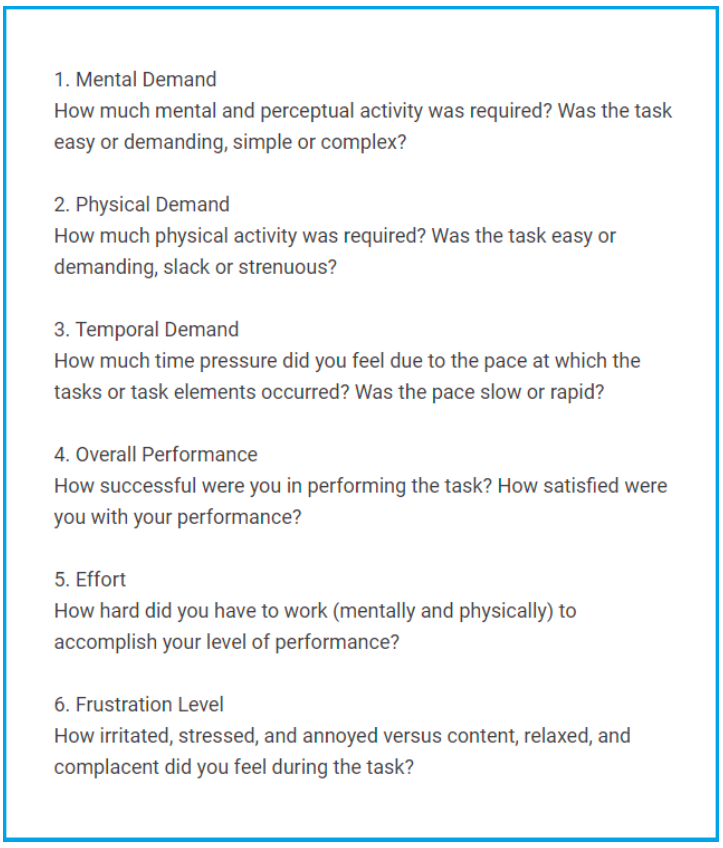


Figure 5.1: The questions of the questionnaire.

3. **SURF Cloud**

SURF Cloud is a server provider that offers servers in various configurations. The Human-AI Collaborative Workflow system is deployed on SURF Cloud. The server configuration is as follows: Ubuntu 20.04 and Docker environment.

5.4. Study Design

The scoring method for task 1 are shown in table 5.2. One answer gets 0 points if contains no key points as ground truth or it is found copied from the entire policy text, 1 points if partial key points are included in the answer, 2 points if all answers are included in the annotator's result but with some invalid information and full points(3 points) if all answers are contained in the annotator's result and without invalid information. Here are some answers from the result set for different scores. The example 1 to example 5 are for the task 1 (Description). The example 6 to example 8 are for the task 3 (Country).

Task ID	Task Name	Ground Truth
1	Description	Policymakers throughout England, UK are hesitant to include school closures in the national lockdown due to learning loss for students that is hard to make up and therefore has potential long-term consequences; as well as much bigger implications for lower socio-economic groups that will widen inequality and reduce social mobility.
2	Type	Closure and Regulations of Schools.
3	Country	UK

Table 5.1: The ground truth for the three tasks.

Example 1 for Task "Description"

This answer got 0 points since the annotator pasted the entire policy text.

- *"Schools and England's second lockdown: further closures would have adverse effects on children and a wider effect on family life. Sandra McNally and Jo Blanden explain how school closures have negative effects on both children and parents, hindering educational outcomes, mental health, work, and general resilience. Policymakers have concluded that the benefits of keeping schools open during the second English lockdown are greater than the risks of increased transmission among school communities. But whether and how much the health risk would reduce is by no means certain. Evidence from Germany indicates that school reopenings there are not associated with an increased risk of infection, but this is in the context of a particular set of disease control measures and is not necessarily applicable in other contexts. During the first national lockdown, the UK government closed all schools in England. The closure resulted in a loss of face-to-face instructional time, which is likely to have significant adverse effects on students' educational outcomes. The evidence finds that inequalities along the lines of family income or type of schooling are very marked. Educators are well aware of the need to make up for the learning losses from the first lockdown, although it is very difficult to do this at the level of intensity, speed, and coverage necessary. Plans for a national tutoring programme are in the process of being rolled out. To enforce school shutdowns again before this has even got started may make a bad situation worse. To the extent that learning loss cannot be made up, this will have long-term effects on students. As the Delve Initiative (2020) puts it 'the skills loss from missing school is not trivial, and is likely to lead to lower earnings, higher risk of poverty and unemployment with impacts on health and life expectancy'.Closing schools again is likely to have big adverse effects on children's learning. Consequently, there were large declines in mental health among those with young children. Overall, school closures have obvious adverse effects on children – which are large in magnitude – and also have a wider effect on family life which further hinders mental health, work, and general resilience. The costs of further closure should not be underestimated. Indeed, a recent commentary in Science states that 'If communities prioritize suppressing viral spread in other social gatherings, then children can go to school.' This is what the government is trying to achieve, and it should continue this path."*

Example 2 for Task "Description"

This answer got 0 points since the result contains no key points.

- *"lockdown due to COVID"*

Example 3 for Task "Description"

This answer got 1 points since the result contains partial key points.

- *"Opening schools during covid has negetive consequences on both children and parents as stated by policy makers. Some children will not having learning equipments for elearning. Pupils will fail to have good grades due to these disturbances"*

Example 4 for Task "Description"

- *"This answer got 2 points since the result contains partial key points but with some invalid information. "School closures have negative effects on both children and parents, hindering educational outcomes, mental health, work, and general resilience. The risks of school closures are high: learning loss for students that is hard to make up and therefore has potential long-term consequences; as well as much bigger implications for lower socio-economic groups that will widen inequality and reduce social mobility. In addition, losing the childcare provided by schools has wider negative effects on parents' work and family life.*
  While all students lose out when schools close, the impact is much greater for some than from others because of huge disparities in the extent to which schools can compensate through the provision of online teaching or how much parents can compensate through home-schooling or other resources.
  Closing schools again is likely to have big adverse effects on children's learning. But this is not the end of the story: many parents struggled with the additional burden of full-time childcare and supervising home learning when schools closed in March.
  Overall, school closures have obvious adverse effects on children – which are large in magnitude – and also have a wider effect on family life which further hinders mental health, work, and general resilience."

Example 5 for Task "Description"

This answer got 3 points since the result contains partial key points and without invalid information.

- *"The risks of school closures are high: learning loss for students that is hard to make up and therefore has potential long-term consequences; as well as much bigger implications for lower socio-economic groups that will widen inequality and reduce social mobility. In addition, losing the childcare provided by schools has wider negative effects on parents' work and family life. All these negatives are set against the possibility of reducing the risk of COVID transmission if schools shut."*

Example 6 for Task "Country"

This answer got 0 points since the result contains no country name at all.

- *"schools shut."*
- *"this will have long-term effects on students."*

Example 7 for Task "Country"

This answer got 1 points since the result contains country name but with some invalid information.

- *"This policy is published in England, in relation to the issues of Schools and England's second lockdown."*
- *"the UK government closed all schools in England"*
- *"LSE Politics and British Politics."*

Example 8 for Task "Country"

This answer got full points(2 points) since the result contains country name and without invalid information.

- *"UK"*
- *"United Kingdom"*
- *"United Kingdom/England"*
- *"UK Government in England"*

- *"This policy was announced in the UK."*

Score	Criteria
0	No answers are contained in the annotator's result.
1	Partial answers are contained in the annotator's result.
2	All answers are contained in the annotator's result, and invalid information.
3	All answers are contained in the annotator's result and no invalid information.

Table 5.2: The scoring rubric for task 'Description'.

Score	Criteria
0	The wrong option is picked.
1	The correct option is picked.

Table 5.3: The scoring rubric for task 'Type'.

Score	Criteria
0	No answers are contained in the annotator's result.
1	The answer is contained in the annotator's result and invalid information.
2	The answer is contained in the annotator's result and no invalid information.

Table 5.4: The scoring rubric for task 'Country'.

5.5. Experimental Conditions

We now have a policy document and AI recommendations of varying quality. The experimental design of this paper is inspired by the study conducted by (Jorge et al. 2019). In this thesis, our goal is to study the impact of the Human-AI Collaborative Workflow system on an annotator's work performance. According to the findings of (Wuand Yuan 2003) and (Gier, Kreiner, and NatzGonzalez 2009), different qualities of highlighting have different effects on the reading time of humans. Taking this as inspiration, we divided the annotators into three groups. A group of people (64 persons) annotating manually, a group of people (64 persons) annotating by the system with accurate advises and a group of people (64 persons) annotating by the system with inaccurate advice. Finally, according to the collected experimental results (completion time and correct rate), we test whether the average work efficiency of experimental group 1 is better than that of the control group and whether the average work efficiency of experimental group 2 is inferior to that of the control group.

1. **Control group:** annotating a policy according to the original manual process.
2. **Experimental group 1:** annotating a policy by the system with accurate advises.
3. **Experimental group 2:** annotating by the system with inaccurate advice.

6

RESULTS AND ANALYSIS

This section collects the experimental results for the three tasks (description, type, and country) from three groups. The three groups include the control group (human), experimental group 1(AI with good advice), and experimental group 2(AI with bad advice).

6.1. Accuracy

It can be seen from Table 6.1, 6.2, 6.3, 6.4, and 6.5, group mean accuracy for the three tasks ('Description', 'Type' and 'Country') were higher for the experimental group 1 (M = 96.88%, SD = 52.61%), (M = 98.44%, SD = 12.5%), and (M = 57.03%, SD = 35.04%) than the control group (M = 49:48%, SD = 83:56%), (M = 90.63%, SD = 29.38%), and (M = 91.41%, SD = 44.43%). Group mean accuracy for the three tasks ('Description', 'Type' and 'Country') were also lower for the experimental group 2 (M = 29.69%, SD = 71.53%), (M = 42.22%, SD = 46.04%), and (M = 32.03%, SD = 68.72%) than the control group (M = 49:48%, SD = 83:56%), (M = 90.63%, SD = 29.38%), and (M = 91.41%, SD = 44.43%). Consistent with the primary hypothesis, AI with accurate recommendation improved annotation task('Description') performance by increasing accuracy, t(126) = 20.11, p < .0001. AI with accurate recommendation improved annotation task('Type') performance by increasing accuracy, t(126) = 6.07, p < .0001. On average, the accurate AI recommendation increases scores for the annotation tasks of 'Description' and 'Type' by 1.43 and 0.07, 95% CI [1.29, 1.57] and 95% CI [0.05, 0.09].

$$accuracy_{Description} = \frac{\sum score_i \cdot number_i}{number_{total} \cdot score_{full}}, (i = 0,1,2,3.score_{full} = 3, number_{total} = 64)(1)$$

Description			
Score	**Control group: Human**	**Exp group 1:AI(Good Advises)**	**Exp group 2:AI(Bad Advises)**
0	11	2	18
1	14	0	37
2	36	0	7
3	3	62	2
M	49.48%	96.88%	29.69%
SD	83.56%	52.61%	71.53%

Table 6.1: Experimental results for task 'Description'.

$$accuracy_{Type} = \frac{\sum score_i \cdot number_i}{number_{total} \cdot score_{full}}, (i = 0,1.score_{full} = 1, number_{total} = 64)(2)$$

Type			
Score	**Control group: Human**	**Exp group 1: AI(Accurate Advises)**	**Exp group 2: AI(Inaccurate Advises)**
0	6	1	19
1	58	63	45
M	90.63%	98.44%	42.22%
SD	29.38%	12.5%	46.04%

Table 6.2: Experimental results for task 'Type'.

$$accuracy_{Country} = \frac{\sum score_i \cdot number_i}{number_{total} \cdot score_{full}}, (i = 0,1,2. score_{full} = 2, number_{total} = 64) \quad (3)$$

Country			
Score	**Control group: Human**	**Exp group 1: AI(Accurate Advises)**	**Exp group 2: AI(Inaccurate Advises)**
0	2	0	29
1	7	55	29
2	55	9	6
M	91.41%	57.03%	32.03%
SD	44.43%	35.04%	68.72%

Table 6.3: Experimental results for task 'Country'.

	P-Value	
Task Name	**Control Group vs. Exp Group 1**	**Control Group vs. Exp Group 2**
Description	<0.0001	<0.0001
Type	<0.0001	<0.0001
Country	<0.0001	<0.0001

Table 6.4: The P-Value for the three tasks.

	t(126)	
Task Name	**Control Group vs. Exp Group 1**	**Control Group vs. Exp Group 2**
Description	20.11	-2.87
Type	6.07	-6.07
Country	-10.13	-7.7

Table 6.5: The results of T-Test for the three tasks.

6.2. Completion Time

According to Table 6.6, 6.7, and 6.8, the experimental group 1 was faster (m:s) on average (M = 6:48, SD = 3:55) than the control group (M = 16:46, SD = 12:16). The experimental group 2 was slower (m:s) on average (M = 27:02, SD = 10:18) than the control group (M = 16:46, SD = 12:16). Consistent with the primary hypothesis, AI with accurate recommendation improved annotation task performance by increasing speed, t(126) = -6.22, p < .0001. AI with inaccurate recommendation decrease annotation task performance by increasing speed, t(126) = 5.14, p < .0001. On average, the accurate AI recommendation increases scores for the annotation tasks of 'Description' and 'Type' by 1.43 and 0.07, 95% CI [1.29, 1.57] and 95% CI [0.05, 0.09]. The inaccurate AI recommendation decreases scores for the annotation tasks of 'Description' and 'Type' by 0.59 and 0.49, 95% CI [-0.99, -0.18] and 95% CI [-0.65, -0.33]

Group Name	Mean	Standard Deviation	Variance
Control Group	16:46	12:16	147:45
Experimental Group 1	6:48	3:55	15:37
Experimantal Group 2	27:02	10:18	106:41

Table 6.6: The ground truth for the three tasks.

	Control Group vs. Exp Group 1	Control Group vs. Exp Group 2
P-Value	<0.0001	<0.0001

Table 6.7: The P-Value of completion time for the three tasks.

	T-Test	
	Control Group vs. Exp Group 1	Control Group vs. Exp Group 2
t(126)	-6.22	5.14

Table 6.8: The T-Test results of completion time for the three tasks.

6.3. User Experience

Tables 6.9, 6.10 and 6.11 show the experimental results of the user experience of the Human-AI workflow in terms of six dimensions for the control group (annotating manually), experimental group 1 (annotating with high-quality AI advice), and experimental group 2 (annotating with low-quality AI advises), respectively. The tables contain survey results of six dimensions related to the system's Mental Demand, Physical Demand, Temporal Demand, Temporal Demand, Temporal Demand, and Frustration. Each row in the table contains the dimension, the sample mean, standard deviation, and variance.

The chart below (6.1) shows the mean values of the control group, experimental group 1, and experimental group 2, respectively. It is clear that the data of experimental group 1 is superior to the other two groups. Experimental group 1 (high-quality AI advice) needs lower mental, physical, and temporal demands than the control group (without AI advice) for about 34.9%, 43% and 12.8%, separately. It is also noticeable that the figures for the control group and experimental group 2, mental demand, performance, and effort, tend to be fairly similar. The number of performance for experimental group 1 is higher than both other two groups. On top of that, it also requires less effort (4.57) and experiences less frustration (4.14) with the support of AI advice of good quality than control group 1 (5.88 and 4.84) and experimental group 2 (5.92 and 5.47).

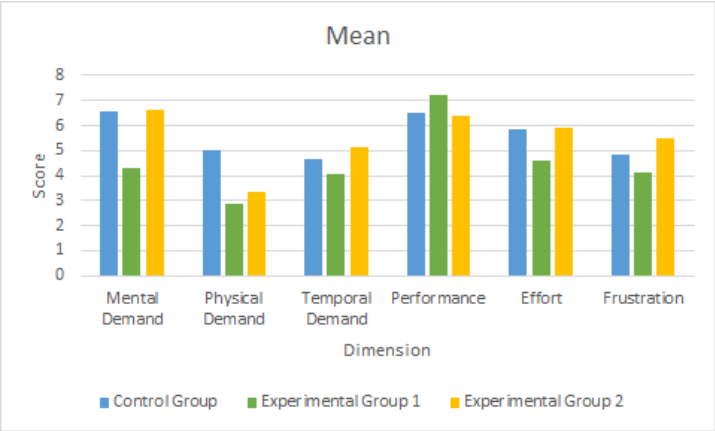


Figure 6.1: The mean value for control group, experimental group 1 and experimental group 2.

Figure 6.2 shows the detailed data for the control group, experimental group 1, and experimental group 2. For the psychological needs, the scores of the control group and the experimental group 2 are clustered around 6 to 8. In contrast, the data of experimental group 1 are clustered around four scores. This shows that the

Human-AI Collaborative Workflow with good advice can reduce the psychological needs of the annotators. A similar situation arises with physical demands, time demands, effort, and frustration.

Figure 6.2: The statistical results for Mental demand, Physical demand, Temporal demand, performance, Effort and Frustration.

Field	Mean	Std Deviation	Variance
Mental Demand	6.59	1.46	2.12
Physical Demand	5	0	0
Temporal Demand	4.68	2.70	7.29
Performance	6.53	1.83	3.35
Effort	5.88	2.24	5.02
Frustration	4.84	2.34	5.49

Table 6.9: The questionnaire results in the perceived workload of the Human-AI Collaborative Workflow (Manual Tasks).

Field	Mean	Std Deviation	Variance
Mental Demand	4.29	1.44	2.06
Physical Demand	2.85	1.74	3.02
Temporal Demand	4.08	2.12	4.49
Performance	7.23	2.82	7.91
Effort	4.57	2.65	7.03
Frustration	4.14	2.93	8.59

Table 6.10: The questionnaire results in the perceived workload of the Human-AI Collaborative Workflow (High quality AI advises).

Field	Mean	Std Deviation	Variance
Mental Demand	6.60	1.96	3.82
Physical Demand	3.32	2.25	5.05
Temporal Demand	5.15	2.41	5.81
Performance	6.37	2.09	4.38
Effort	5.92	2.28	5.19
Frustration	5.47	2.51	6.29

Table 6.11: The questionnaire results in the perceived workload of the Human-AI Collaborative Workflow (Low quality AI advises).

7

DISCUSSION AND LIMITATION

In this experiment, experts must translate complex tasks and then score the annotation results according to the scoring criteria. For example, in task 1, to generate an article summary, an expert needs to summarize the critical points of a document first and then check how many critical points each annotation result contains and whether it contains invalid information. For task 2 (multiple-choice), select "policy type." Experts need to make the ground truth first, and then the system can directly judge right or wrong (this step no longer requires expert participation).

It can be seen from the experimental results in Table 6.1, 6.2, and 6.3 that the accuracy of the experimental group 1 (96.88%) is much higher than that of the control group. The AI model of experimental group 1 gave a 3-point answer as a suggestion. Almost all the annotators in experimental group 1 gave correct answers based on the system's suggestions. 62 of the 64 answers in total are entirely in line with the AI's recommendations, and 2 of the two answers are shown as 'null'. For the control group, the low accuracy was not due to too many wrong answers. In fact, most of the annotators in this group gave 2-point answers, while the full-point answer was only 3 points. This resulted in lower accuracy in the control group.

Table 6.6 shows that the average time for the control group, experimental group 1, and experimental group 2 was 16 minutes 46 seconds, 6 minutes 48 seconds, and 27 minutes 02 seconds, respectively. Experimental group 1 is only one-quarter of the length of the control group. This shows that good AI suggestions can significantly improve annotators' efficiency. This may be because annotators tend to trust AI's recommendations. When the AI model tells the annotator that the current recommendation has high confidence, the annotator tends to choose to spend less time checking whether the answer is correct or save the AI-recommended answer. On the other hand, experimental group 2 had more average time than the control group. This situation also makes sense because when annotators find AI recommendations with low confidence, in addition to spending time looking for answers, they also spend time reading AI recommendations and judging whether they should be discarded entirely, or modify it based on it, or use this answer directly. Therefore, it can be concluded that the quality of AI suggestions is positively correlated with the efficiency of the annotators.

According to the results of user experiment, it is reasonable to infer that a Human-AI Collaborative Workflow with good AI suggestions can increase the productivity of annotators, while bad AI suggestions can have the opposite effect. This can be confirmed from the bar chart of Performance in the second row and second column of Figure 6.2. It shows that most annotators gave scores higher than 6, and nearly half gave full scores.

Due to limited research funding, annotation tasks and questionnaires must be completed in approximately 30 minutes. Therefore, the number of topics is small, and the form is single. More open-ended questions can be added to this questionnaire with more time and money, such as asking the annotators if they have any other comments about the system. There are more questions about user experience. For example, design separate questions for the three tasks, specifically for good and bad suggestions, and ask more detailed questions for different AI suggestions (highlighting, confidence labels, etc.).

The Human-AI Collaborative Workflow system can also be further optimized. For example, Task 1 (Summary) can configure critical points (e.g., who, when, what, etc.) so that it can be split into multiple answer boxes. Then, the system divides the annotation results into multiple sub-answers based on crucial points and presents them to experts. Instructions for Task 3 can also be improved by asking the annotators to write common country names rather than abbreviations or other terms. These improvements can increase the efficiency with which experts review annotation results.

8

CONCLUSION AND FUTURE WORK

This thesis measures the impact of the Human-AI Collaborative Workflow approach on the accuracy and completion time of annotation tasks by having annotators perform a series of annotation tasks on policy documents using the Human-AI Collaborative Workflow System. In addition, we separately measured the impact of AI midway products (e.g., highlighted answers, confidence in multiple-choice suggested options, confidence in fill-in-the-blank suggested solutions) on non-expert work outcomes (accuracy and completion time). The experimental results of this study show that AI suggestions directly impact the annotator's accuracy and completion time during the annotating process. Furthermore, the quality of AI recommendations was positively correlated with the accuracy of the annotation results and completion time. More specifically, high-quality AI suggestions can improve the accuracy of annotating results and significantly reduce the time-consuming annotating work (a quarter of the manual annotating time). However, low-quality AI suggestions reduce accuracy and cause annotators to spend more time deciding whether or how to use AI suggestions.

The results often contain redundant or invalid information when using QA models to generate fill-in-the-blank tasks. This study initially planned to use knowledge graphs to assist QA models in improving the quality of AI recommendations. However, due to time constraints, knowledge graphs cannot be integrated into the Human-AI Collaborative Workflow system. This is an essential direction for improving Human-AI Collaborative Workflow in the future.

Bibliography

[1] Alan Aipe and Ujwal Gadiraju. Similarhits: Revealing the role of task similarity in microtask crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media*, pages 115–122. 2018.

[2] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3665–3674, 2014.

[3] Joshi-C. Nolan L. Henstra-Hill R. Shaw L. Hook S Benedikt, L. Human-in-the-loop ai in government: a case study. *Proceedings of the 25th International Conference on Intelligent User Interfa-ces (pp. 488-497)*, 2020.

[4] Anna CS Bodén, Jesper Molin, Stina Garvin, Rebecca A West, Claes Lundström, and Darren Treanor. The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice. *Histopathology*, 79(2):210–218, 2021.

[5] Alon Brutzkus and Amir Globerson. Why do larger models generalize better? a theoretical perspective via the xor problem. In *International Conference on Machine Learning*, pages 822–830. PMLR, 2019.

[6] Sikai Chen, Yue Leng, and Samuel Labi. A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information. *Computer-Aided Civil and Infrastructure Engineering*, 35(4):305–321, 2020.

[7] Davis-L. Ippolito D. Reif-E. Yuan A Coenen, A. Wordcraft: a human-ai collaborative editor for story writing. 2021.

[8] Jameel-S. Schneider D. Paredes-H. Fonseca B Correia, A. A workflow-based methodological framework for hybrid human-ai enabled scientometrics. *IEEE International Conference on Big Data (Big Data) (pp. 2876-2883)*, 2020.

[9] Lorrie F Cranor. A framework for reasoning about the human in the loop. 2008.

[10] J. A Crowder. Computing with words: Human-ai collaboration. 2012.

[11] Michelangelo Diligenti, Soumali Roychowdhury, and Marco Gori. Integrating prior knowledge into deep learning. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 920–923. IEEE, 2017.

[12] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[13] Zhang-H. Chen Y. Li F Dong, Z. Interpretable drug synergy prediction with graph neural networks for hu-man-ai collaboration in healthcare. 2021.

[14] Yucong Duan, Lixu Shao, Gongzhu Hu, Zhangbing Zhou, Quan Zou, and Zhaoxin Lin. Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph. In *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 327–332. IEEE, 2017.

[15] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with ai systems. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[16] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.

[17] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[18] Ujwal Gadiraju and Jie Yang. What can crowd computing do for the next generation of ai systems? In *CSW@ NeurIPS*, pages 7–13, 2020.

[19] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–29, 2017.

[20] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and human-centered experiments*, pages 6–26. Springer, 2017.

[21] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 5–14, 2017.

[22] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. *Computer Supported Cooperative Work (CSCW)*, 28(5):815–841, 2019.

[23] Vicki Silvers Gier, David S Kreiner, and Amelia Natz-Gonzalez. Harmful effects of preexisting inappropriate highlighting on reading comprehension and metacognitive accuracy. *The Journal of general psychology*, 136(3):287–302, 2009.

[24] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 241–249, 2020.

[25] Gabriel Hartmann, Zvi Shiller, and Amos Azaria. Deep reinforcement learning for time optimal velocity control using prior knowledge. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 186–193. IEEE, 2019.

[26] Cheng Y. T. Chen L. L. Hsu J. Y. J Huang, Y. C. Human-ai co-learning for data-driven ai. 2019.

[27] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[28] Doris Jung-Lin Lee and Stephen Macke. A human-in-the-loop perspective on automl: Milestones and the road ahead. *IEEE Data Engineering Bulletin*, 2020.

[29] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[30] F Li. Reproducible prognostic and health management for complex industrial system using human-ai collaboration. *Doctoral dissertation, University of Cincinnati*, 2021.

[31] Yancong Lin, Silvia L Pintea, and Jan C van Gemert. Deep hough-transform line priors. In *European Conference on Computer Vision*, pages 323–340. Springer, 2020.

[32] Wilson A. Gunning D Liu, J. Workflow-based human-in-the-loop data analytics. *Workshop on Human Centered Big Data Re-search (pp. 49-52)*, 2014.

[33] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[34] Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 144–152, 2019.

[35] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, 2018.

[36] Bevilacqua F Scurto, H. Appropriating music computing practices through human-ai collaboration. *Journées d'Informatique Musicale*, 2018.

[37] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, pages 882–892, 2022.

[38] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, pages 77–87, 2021.

[39] Jen-Her Wu and Yufei Yuan. Improving searching and reading performance: the effect of highlighting and text color coding. *Information & Management*, 40(7):617–637, 2003.

[40] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In *Proceedings of the second workshop on data management for end-to-end machine learning*, pages 1–4, 2018.

APPENDIX A

Instructions Task 1:

In this annotation task, you need to summarize a policy. You should first read the policy on the right side and check if the highlighted part is what you want and do some modifications. Finally, click the 'Save' button to save your answer to the database. (Note: your answer should contain when and who did what policy at least.)

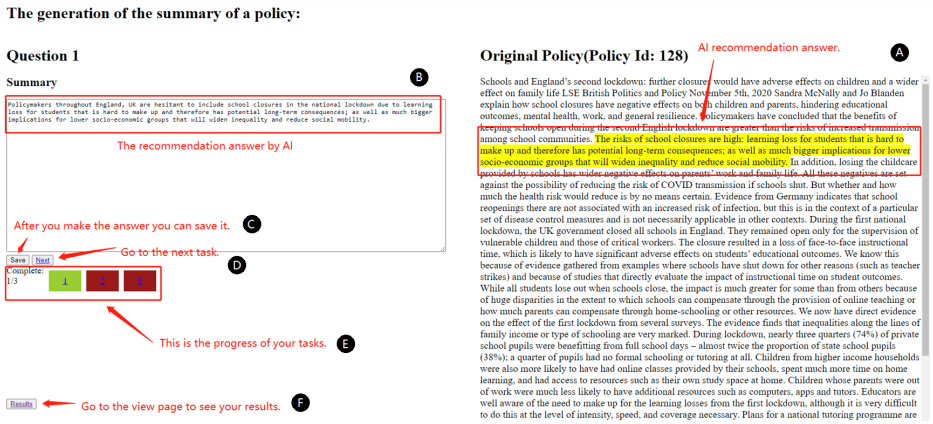


Figure 8.1: Task 1 description in the instructions.

Instructions Task 2:

In this task, you need to do a selection. You need to read the policy text. And then, read the question and the clarification of the property on the left side. Finally, based on the highlight and AI recommendation option, select an answer from the options and click the 'Save' button. Click the 'Next' button to go to the next task. (Note: the AI Recommendation label can be red or green, which means that it is low and high confidence respectively. You should pay more attention when it is red.)

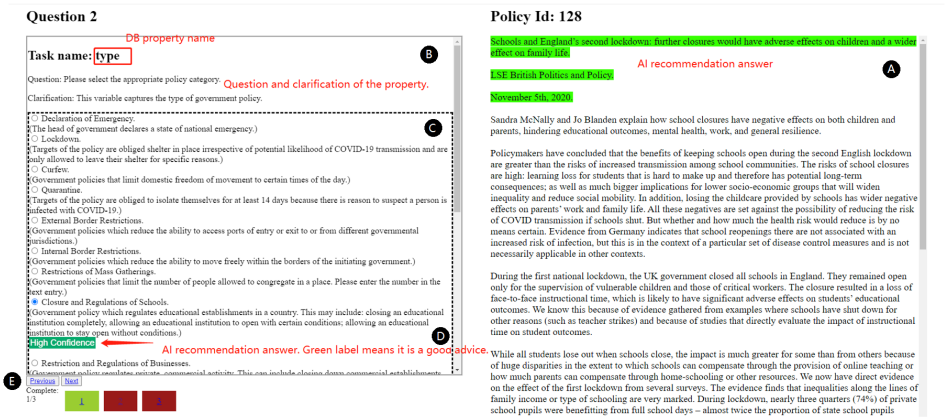


Figure 8.2: Task 2 description in the instructions.

Instructions Task 3:

This is a fill-in task, the same as the previous tasks. You need to read the policy text and the questions and the clarification of the property. And based on one of the candidate answers, you fill the answer in

the text box. Click the ‘Save’ button. Now, all four tasks are done. (Note: when you click the radio button of the AI recommendations, the system will highlight the corresponding answer on the right side and fill the recommendation in the text box.)

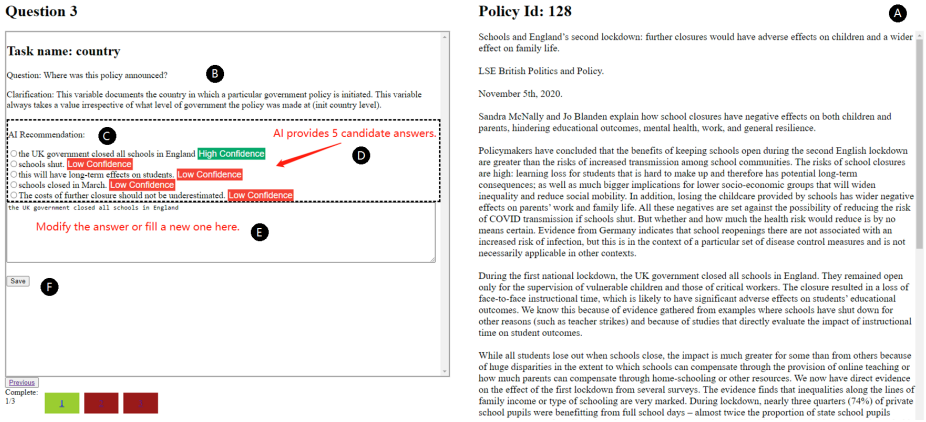


Figure 8.3: Task 3 description in the instructions.

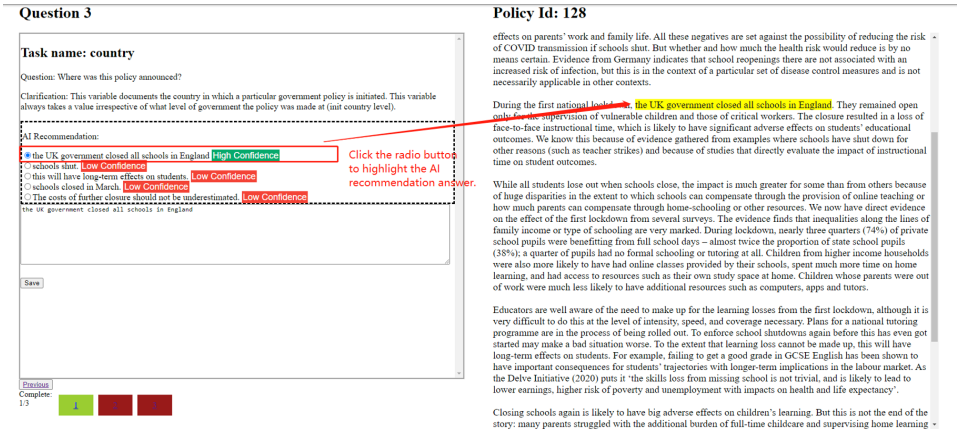


Figure 8.4: Task 3 description in the instructions.

APPENDIX B

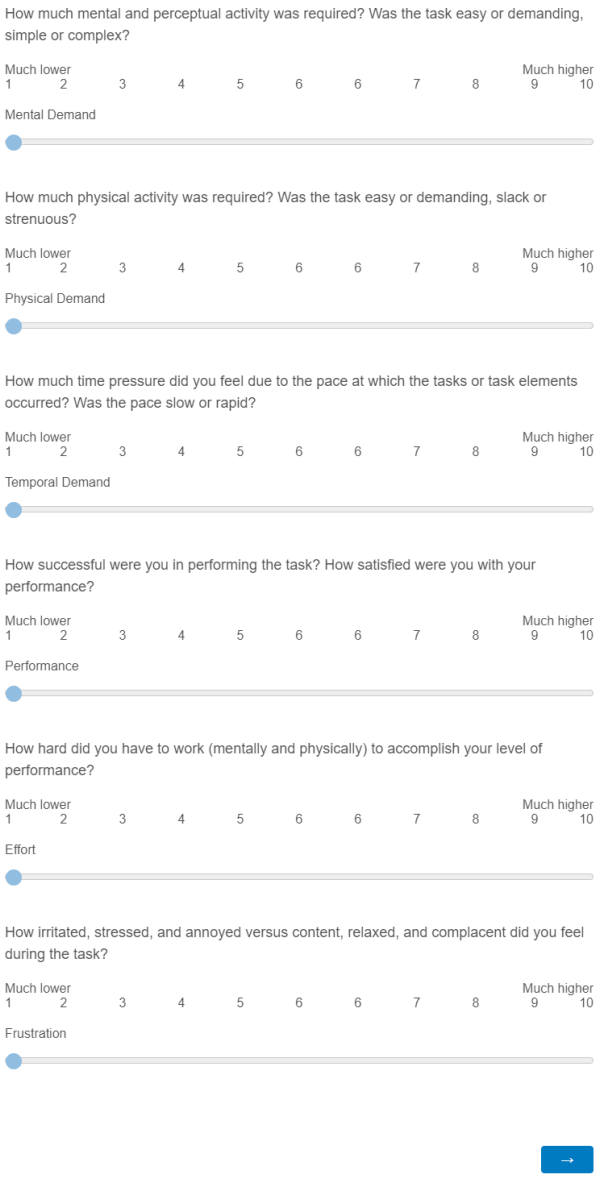


Figure 8.5