5 APPENDIX

5.1 Usability Study

5.1.1 Experimental Setup. To test AnnoRank in a real environment we conducted a small-scale usability study. In total, we had 14 participants with a computer science, management, and physics educational background. Each assignment was tested by 5 participants. Following [1] we conducted interviews with the participants to evaluate each of the above-presented use cases in terms of appearance, content, navigation, and functionality. Additionally, we evaluated the technical setup of AnnoRank. The technical setup of AnnoRank was evaluated by 5 participants out of which 1 had Windows OS and 4 Mac OS.

5.1.2 Amazon Dataset- Interaction Annotate UI. **Task Description:** The participants were asked to select the top-3 items that best matched the shopping category and product description while prioritizing the items based on a combination of price and rating. Results: The user interface (UI) was assessed as intuitive, simple, and practical. Participants found the task easy to comprehend and the annotation process straightforward to follow. We received the suggestion of making the Task Description Field and the Query Field stick at the top of the page while scrolling down through the list of items. A participant suggests to add pictures to the items to ease the annotation task. AnnoRank supports adding pictures as both items and queries. Another participant suggested adding functionality such that terms from the query that appear in the items are highlighted. No technical problems were encountered during the experiment.

5.1.3 Recruitment Dataset - Interaction Annotate UI. **Task Description:** The participants were asked to select the top-3 candidates given the job requirements. **Results:** The UI was evaluated as usable and intuitive. The participants found the task to be easy to understand but harder to complete due to the lack of expertise. To ease the annotation task the participants suggested adding a functionality such that terms from the query that appear in the items are highlighted. Similarly, participants suggested having the Task Description Field and the Query Field stick at the top of the page while scrolling down through the list of items. One participant suggested making the View Button more visible. No technical problems were encountered during the experiment.

5.1.4 Recruitment Dataset - Score Annotate UI. **Task Description:** The participants were asked to annotate the job candidate's education, work experience, and skills considering job requirements with a score from 1 to 5. **Results:** The UI was evaluated as intuitive. Participants found the task to be easy to follow; however, assessing the candidate's skills was considered harder given the lack of expertise and the long list of skills. To ease the annotation task, the participants suggested adding functionalities such that terms from the query that appear in the items are highlighted, and similar positioning between features asked in the query field and on the candidate's list field would be better. Similarly, participants suggested having the Task Description Field and the Query Field stick at the top of the page while scrolling down through the list of items. No technical problems were encountered during the experiment.

5.1.5 Flickr Dataset - Score Annotate UI. **Task Description:** The participants were asked to annotate if the displayed caption is 1-relevant or 0-not relevant for the displayed image. **Results:** The UI was evaluated as intuitive and straightforward by all participants. All participants found the task easy to follow and easy to complete. We received the suggestion of displaying the Task Description before the assessment started to offer the opportunity to display a longer task description. No technical problems were encountered during the experiment.

5.1.6 Technical Setup. **Task Description:** Participants were asked to download the AnnoRank app and the instructions presented in the ReadMe file to be able to start the app. **Results:** The set-up was evaluated by the 5 participants to be straightforward and easy to follow. The install time was considered to be reasonable. Installing the app on Windows required extra requirements that we added in the README.md after conducting the usability study. Unfortunately, one participant using Mac OS 11.5 was unable to install Docker as their older system was not supported.

5.1.7 Conclusion. To test AnnoRank in a real environment we conducted a small-scale usability. Each assignment was tested by 5 participants. AnnoRank was evaluated in terms of appearance, content, navigation, functionality and technical setup. 9 out of 11 reported that AnnoRank has an intuitive and usable UI. Regarding the annotation assignments, participants reported that the task was easy to follow and straightforward to complete. During the interaction with AnnoRank no problems or errors were reported. The set-up was evaluated by most participants to be straightforward to follow.

According to the feedback received during the usability study, we performed the following improvements: made the View Button more visible, and added the possibility to automatically highlight terms from the query in the item's text to ease the annotation task.