|  |  |  |
| --- | --- | --- |
| **Paper** | **Human Evaluation Size** | **Notes** |
| COMET-ATOMIC-2020 | 3k random instances per KG. Three different knowledge graphs. | “To ensure high-quality annotations, we qualified a pool of 173 workers through a paid qualification task that tested their ability to follow directions and provide reasonable answers to the qualification test. The qualification test contained 6 manually selected tuples from ATOMIC and CONCEPTNET, including both easy and tricky relations to annotate. A worker was qualified if they provided 100% acceptable answers. Workers providing 5 of 6 correct answers were also accepted only when they provided a reason- able written substantiation for their incorrect choice.”  “Workers were presented with knowledge tuples in the form of (*head*, *relation*, *tail*) for an- notation. To expedite the human assessment of the tuples, each *relation* (e.g., xWant or AtLocation) was translated into a human-friendly natural language form (e.g., “as a result, PersonX wants” and “located or found at/in/on”, respectively; cf. Appendix ). The workers were asked to rate the tuples along a 4-point Likert scale: *always/often* – the knowledge assertion presented is always or often true, *some- times/likely* – it is sometimes or likely true, *farfetched/never* – it is false or farfetched at best, and *invalid* – it is invalid or makes no sense. Any tuples receiving the former two labels are ranked as Accept and latter two as Reject. The workers were also given a choice to opt out of assessment if the concepts were too unfamiliar for a fair evaluation (No Judgment). Each task (HIT) included 5 tuples of the same relation type, and each tuple was labeled by 3 workers. For the results, we take the majority vote among the 3 workers.” |
| CommonsenseEmergence-2021 | ? | “For the human evaluation (Accept % in Table 3), we use the procedure described in [Hwang et al., 2021].” |
| Commonsense QA 2.0 | In order of 1000s?! |  |
| Chain-of-Thought | Depends on the dataset:   * SayCan: 120 * GSM8k: 1,319 * SVAMP: 1000 * AQuA: 254 * MAWPS: 395-600 | * Standard deviation is reported for multiple runs with different random seeds. |
| I’m Not Mad | * Sample 100 events from ATOMIC. 5 evaluators each. Each event has 9 relations and 10 options. So total of 9,000 samples. |  |
| It's Not Rocket Science | Two types of evaluations:   * Absolute evaluations: 50 random narratives for each task. Two tasks of Idioms and Similes. * Comparative evaluation: 100 random narratives for idioms and 75 for similes. |  |
| Penguins Don’t Fly | I guess 1,500. As there are 300 generics and 5 top from each generic are annotated. |  |
| Symbolic Knowledge Distillation | * 3,000 before filtering, so at least 1,000 are there. |  |