Rillor

RESEARCH REPORT

Language models as annotators: agreement with people in published studies

What published studies find when an LLM is used as an annotator: agreement with crowd workers and experts, task by task, and how to check model labels.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Published studies find that a language model used as an annotator can match or beat crowd workers on some short-text tasks and fall short on many expert, long-text and subjective tasks, with mixed results on span tasks, so each task needs its own check against human labels [3, 7, 12, 13, 14, 17].

- Results depend on the task. In a 2023 study of 6,183 tweets and news articles, a commercial chat model given instructions and no examples was about 25 percentage points more accurate than crowd workers on average [3]. A replication of 27 social science annotation tasks the same year found nine where model labels had precision or recall below 0.5 [7].
- Expert labeling shows a gap between models and experts. In a 2025 study of five finance, law and biomedicine datasets, four instruction-tuned models matched expert consensus labels on 64.2% to 68.7% of items on average, and reasoning models were not significantly more accurate in most settings [12].
- On subjective labels, agreement depends on which people. Most of nine models tested on politeness and offensiveness matched White participants' labels more closely than Asian or Black participants' [17], and a model's errors in coding refugee interviews were not random with respect to who was interviewed [18].
- Agreeing with itself does not make a model right. On average, two runs of one commercial chat model gave the same label to 91% of items at temperature 1 and 97% at temperature 0.2 [3]. The authors of a 2024 evaluation of 13 models argue that ensembling does not repair wrong labels, because models show high internal agreement even when inaccurate [6].
- Agreement with people is not the same as being right. In a 2026 preprint, model codes overlapped less with human codes (mean Jaccard 0.30) than human codes did with each other (0.52), yet a blind expert preferred the two at statistically indistinguishable rates [26].
- Model labels used as data can mislead. Treating them as true labels produced substantial bias and invalid confidence intervals even at 80% to 90% accuracy [22], and replications of 37 annotation tasks reached incorrect conclusions on about 31% of hypotheses with state-of-the-art models [21]. Methods that combine model labels with a smaller set of human labels give valid estimates [22, 23, 24].
- People shown a model's suggestions adopt them. In a 2025 experiment, the share of crowd labels matching the suggestions rose from 40% without assistance to 81% to 87% with it, and the model's measured F1 against those labels rose by about 0.32 [30].

Background

What annotation is for

Annotation attaches labels to items: a class for a post, the spans of a translation error, a code for an interview passage, a choice between two responses [3, 13, 18, 32]. The labels train models, test them and feed downstream statistical analyses [7, 21, 22]. A 2024 survey divides the use of language models in this work into generating annotations, assessing them and using them [34].

Before models, labels came from trained annotators, such as research assistants or domain experts working from written guidelines, and from crowd workers on paid platforms [3, 4, 12]. Plank argues that annotation projects assume one ground truth and neglect genuine variation in human labeling, from disagreement, subjectivity or several plausible answers, and that this variation affects data, modeling and evaluation [19].

How a model becomes an annotator

The studies set a model up as an annotator in five ways:

- **Zero-shot:** the model gets the instructions and the item, with no labeled examples [3, 9].
- **Few-shot:** a few labeled examples go into the prompt [9, 11].
- **Explain-then-annotate:** the model first writes explanations for labeled examples, and those explanations form a few-shot chain-of-thought prompt [11].
- **Fine-tuning:** an open-weight model is trained on some human labels [10].
- **Assistance:** the model's suggestion is shown to a human annotator [30, 31].

A related use labels preferences: a model chooses the better of two responses, and those labels stand in for human preferences in reinforcement learning from feedback [32].

What "agreement with people" means

The studies use four kinds of measure:

- **Accuracy against a reference:** labels taken as correct, from trained annotators [3], expert consensus [12] or a known fact, such as a politician's party [4].
- **Agreement with annotators:** kappa [6, 15], a gamma score for spans [13] or Jaccard overlap for sets of codes [26], with the model treated as one more coder.
- **Consistency with itself:** the share of matching labels when the same model labels the same items again, sometimes reported as Krippendorff's alpha [3, 4, 8].
- **Preference alignment:** the share of a model's choices between two responses that match human choices [32].

These numbers do not convert into one another. Rao and Callison-Burch show that the same verdicts can support widely different agreement figures depending on the scale, the cases kept, the handling of abstentions and invalid outputs, and how verdicts are pooled [27].

Method

What was examined. Studies from 2021 to 2026 that compare model labels with human labels, and papers on what happens when model labels train classifiers or feed estimates.

Sources. 34 papers: cost and training-data studies [1, 2, 33], comparisons with crowd workers, trained annotators and classifiers on social science text [3, 4, 5, 6, 7, 9, 10, 11], expert, span, long-text and non-English tasks [12, 13, 14, 15], subjective tasks [16, 17, 18, 19], consistency and model change [8, 20], downstream statistics [21, 22, 23, 24, 25], agreement and reporting [26, 27, 28], human review and crowd work [29, 30, 31], preference labels [32] and a survey [34]. Preprints are marked in Sources. Every source was opened on 8 November 2026.

How. Figures come from each paper's full text where its tables and results were read, and from the abstract otherwise. Each figure stays with its task, its model family and its year. Models are described by family, and product names appear only in cited titles. Results are not averaged, because the studies use accuracy, F1, kappa, alpha, gamma and Jaccard on different tasks, and a search for systematic reviews and meta-analyses found none that pools them. Costs are the prices each study reports for its own year.

Findings

1. Early cost studies: model labels as training data

The first studies asked whether model labels could replace human labels in a training set, judged by how well a model trained on them performed.

Wang and colleagues (2021) found that, to bring a downstream model to the same performance on a range of language understanding and generation tasks, labels from a commercial language model cost 50% to 96% less than labels from people [1]. Combining the model's pseudo labels with human labels did better still on a limited labeling budget [1].

Ding and colleagues (2023) tested the idea on SST-2 sentiment [2]:

Training labels	Cost (USD)	Downstream accuracy
3,000 human	221 to 300	88.47
3,000 from the model	11.31	87.75
6,000 from the model	22.63	89.29

Source: [2], Table 1 and Results.

Downstream accuracy is that of a model fine-tuned on each label set and tested on human-labeled test data, and the human labeling time leaves out preparing the instructions [2]. 6,000 model labels beat 3,000 human labels for about 10% of the cost [2]. Both studies measure the value of labels as training data.

2. Comparisons with crowd workers on short social media and news text

Gilardi, Alizadeh and Kubli (2023) set a commercial chat model against crowd workers on four samples of tweets and news articles (n = 6,183) for relevance, stance, topics and frame detection, with accuracy measured against the authors' trained annotators [3]. The crowd workers were high-qualification workers on a paid platform, with approval rates over 90%, located in the United States [3].

- Zero-shot accuracy exceeded the crowd workers' by about 25 percentage points on average [3].
- On relevance, accuracy was 70% for content moderation tweets, 81% for content moderation news articles, 83% for US Congress tweets and 59% for 2023 content moderation tweets [3].
- Agreement between two coders of the same kind was about 56% for crowd workers, 79% for trained annotators, 91% for two model runs at temperature 1 and 97% at temperature 0.2 [3]. The model figures compare two runs of the model with each other.

- The cost per annotation was under $0.003, about thirty times cheaper than the crowd platform [3].

Törnberg (2023) asked a later commercial model to classify the party of US politicians from 500 of their tweets during the 2020 election, a task with a known answer [4]. Ten crowd workers labeled each message and two political science researchers labeled all of them [4]. The model outperformed every individual human classifier and the combined crowd, had much higher reliability, higher at lower temperature, and showed no significantly different bias from the experts, while the crowd workers were significantly more biased toward guessing Democrat [4].

AnnoLLM (2024) tested explain-then-annotate against human references [11]:

Task	Human reference	Zero-shot	8-shot	Explain-then-annotate
Query-keyword relevance, test accuracy	71.5 (crowd annotators)	70.00	67.80	75.60
Word sense (WiC), test accuracy	80.0 (from the dataset's authors)	59.79	66.36	69.17

Source: [11], Tables 2 and 3.

On BoolQ, where three of the authors reached 89% on 110 examples, the method surpassed the human annotators [11]. On word sense the authors report "a considerable disparity" with crowd annotators [11]. One method gave three outcomes on three tasks.

3. Agreement varies by task and by label

Zhu and colleagues (2023) relabeled five social computing datasets with a commercial chat model and scored it against the original human labels [5]:

Task	Accuracy
Stance detection	0.612
Hate speech	0.571
Sentiment	0.649
Bot detection	0.639

Task	Accuracy
Russo-Ukrainian sentiment	0.573

Source: [5], Table 3.

Average accuracy was 0.609 [5]. Inside one task, precision was 0.353 on tweets containing anti-Asian hate speech and 0.791 on tweets containing counterarguments to it [5].

Ziems and colleagues (2024) tested 13 models zero-shot on 25 English computational social science benchmarks [6]. On classification the models did not outperform the best fine-tuned models but reached fair agreement with humans [6]. The top misinformation model reached κ = 0.55 with gold labels, above the κ = 0.51 inter-human agreement an earlier study reported [6]. On 8 of 17 tasks agreement was moderate to good, κ = 0.40 to 0.65, and on 12 of 17 fair to strong [6]. Even the strongest models were unusably low on some tasks, and the authors suggest using a model as one of several labelers where it matches human agreement [6].

Pangakis, Wolken and Fasching (2023) replicated 27 annotation tasks across 11 datasets from social science articles, over 200,000 text samples [7]. Median accuracy was 0.850 and median F1 0.707, with F1 from 0.059 to 0.969 [7]. Nine tasks had precision or recall below 0.5, and three had both below 0.5 [7]. Performance was "highly contingent on both the dataset and the type of annotation task", so the authors call for validation against human labels task by task [7].

Kristensen-McLachlan and colleagues (2025 version) ran two binary tasks, political content and exemplar tweets, on 1,000 tweets from US news media each, with zero- and few-shot, generic and custom prompts, on open-source models and a commercial chat model [9]. A supervised classifier generally outperformed both, and the commercial model's recall fell as low as 0.1 on exemplar tweets with a zero-shot generic prompt [9].

Alizadeh and colleagues (2024 version) found that fine-tuned open-weight models matched or surpassed commercial models used zero-shot, though they lagged a fine-tuned commercial model, and that fine-tuning was preferable to few-shot training with a modest quantity of annotated text [10].

4. Expert, span and long-text tasks

Expert labels. Tseng and colleagues (2025) built five datasets in finance, law and biomedicine, 200 instances each, with documented guidelines and expert consensus labels, and tested four non-reasoning and two reasoning models [12]. With plain prompting the four instruction-tuned models averaged 64.2%, 65.3%, 68.7% and 67.8% accuracy [12]. Scores ranged from 83.5% to 86.5% on the law dataset CUAD and from 43.0% to 47.0% on FoDS [12]. Chain-of-thought and self-consistency brought "only marginal or even negative performance gains", reasoning models showed no statistically significant improvement in most settings, and the authors describe "a notable gap between current LLMs and human experts" [12].

Spans. Kasner and colleagues (2026 version) compared models and people on three span tasks, errors in text generated from data, translation errors and propaganda techniques, with over 40,000 model and human span annotations [13]. Models had only moderate agreement with human annotators but made errors at a similar rate to skilled crowd workers [13]. On propaganda spans, expert annotators before consolidation reached γ = 0.31 in an earlier study, as cited, and the top model γ = 0.16 [13]. On the data-to-text task three models mostly reached or surpassed human agreement [13]. Crowd annotation of 1,000 outputs cost about $500, and the same work with a reasoning model $3.60 [13].

Long ethnographic texts. Goodall and colleagues (2026 preprint) asked seven models to code 121 ritual features across 567 ethnographic excerpts [14]. Mean F1 on the larger dataset ranged from 0.12 to 0.41 across models and prompting conditions, and every model beat a majority-class baseline, whose mean F1 was 0.02 [14]. On a second dataset two human coders reached a mean Cohen's kappa of 0.57, with κ = 0.92 for singing and κ = 0.25 for generic movement [14]. Features people found hard to agree on were hard for models too, with human kappa and model F1 correlated at r = 0.61, yet even where people agreed, models fell short [14].

A non-English task. Zhao and colleagues (2026 preprint) compared four models with trained annotators on story grammar elements in spoken Mandarin narratives [15]. The best-performing model reached κ = 0.794 with human raters against 0.872 between humans and cut annotation time by 65%, while a locally deployable lightweight model did substantially worse [15]. Reliability fell on young adult narratives, and the authors call for continued human oversight in semantically complex tasks [15].

5. Subjective tasks: whose labels the model matches

On subjective tasks the reference label is itself a choice among people [19].

Sun and colleagues (2025 version) tested nine models on politeness and offensiveness ratings from the POPQUORN dataset [17]. Zero-shot, most models' predictions for both tasks aligned more closely with labels from White participants than from Asian or Black participants, with a minor gender bias favoring women on politeness [17]. Telling the model to answer as a member of a group did not consistently improve the match and sometimes made it worse [17].

Ashwin, Chhabra and Rao (2023) coded interview transcripts of Rohingya refugees in Cox's Bazar, Bangladesh [18]. The models' errors were not random with respect to the characteristics of the people interviewed, and simpler supervised models trained on high-quality human annotations gave less measurement error and bias [18].

Felkner, Thompson and May (2024) asked a commercial chat model to extract stereotypes from community survey responses to build bias benchmarks [16]. Less than half of the extracted predicates were classified correct [16]. Over thirty percent of its predicates for WinoQueer and nearly twenty percent for WinoSemitism expressed a stereotype or bias not present in the response, and negations conflated victims with

perpetrators, most severely for lesbian respondents, where nearly 30% of the model's outputs did so [16]. Bias scores on the human-built and model-built benchmarks correlated at $R^2 = 0.14$ for WinoQueer and 0.73 for WinoSemitism [16].

6. Consistency across runs, prompts and model versions

Reiss (2023) labeled 234 website texts as news or not news, repeating every configuration 10 times for 46,800 inputs [8]:

Comparison	Krippendorff's alpha
Two temperature settings, single runs	0.75
Two temperature settings, majority of ten runs	0.91
Two instructions differing only in "classify" and "rate"	0.6 or below
All 45 pairs of instructions	0.24 to 0.7, mean 0.43

Source: [8].

The 0.75 fails the reliability level Krippendorff recommends, as Reiss cites it, and Reiss advises against using the model for annotation without supervision [8]. Gilardi's 91% and 97% at two temperatures [3] and Törnberg's higher reliability at lower temperature [4] point the same way.

Hosted models also change under the same name. Comparing March and June 2023 versions of two hosted models, Chen, Zaharia and Zou found one fell from 84% to 51% accuracy on telling prime from composite numbers [20]. It is a study outside annotation, and it shows that the "same" service can change substantially in a short time [20].

7. Agreement is not the same as being right

Liu and colleagues (2026 preprint) had five model systems and three trained human coders apply a 72-item hierarchical codebook to 2,560 educator messages from a school education platform [26]. A domain expert then judged 855 pairwise comparisons of code sets without knowing their source [26]. Agreement between model and human coding (mean Jaccard 0.30) was well below agreement between human coders (0.52), yet the expert preferred human and model coding at indistinguishable rates, 51.5% against 48.5% ($p = 0.537$) [26]. For several codes, human consensus encoded a shared bias that the expert rejected in favor of the model's reading, and the authors conclude that agreement-based evaluation is insufficient for automation decisions [26].

Rao and Callison-Burch (2026 preprint), writing about model judges, show that for non-degenerate binary verdicts Pearson's r, Spearman's ρ, Kendall's tau-b, the phi coefficient and the Matthews correlation coefficient are exactly the same statistic, and Cohen's κ differs only through a marginal-mismatch factor [27]. Protocol choices alone moved reported accuracy from 0.551 to 0.899 and carried κ across zero without changing a single verdict [27].

Kunilovskaya and colleagues (2026) audited 2,667 annotation tasks in 1,603 papers from ACL venues, 2018 to 2025, with a model-assisted extraction pipeline checked against a human-adjudicated sample [28]. Papers often omitted details needed to assess annotation validity, among them training, language proficiency, compensation, socio-demographics, adjudication and agreement values, especially in model-evaluation studies [28].

Kasner and colleagues also checked correctness by hand. In 216 samples, 49.5% of model annotations and 50.5% of the model's reasons were correct and 9.2% of annotations partly correct, and models confused related categories and used a less diverse spread of categories than human annotators [13].

8. Using model labels downstream

Estimates. Egami and colleagues (2023) show that using model labels directly in downstream statistical analyses "leads to substantial bias and invalid confidence intervals, even with high surrogate accuracy of 80-90%" [22]. Their design-based supervised learning estimator combines model labels with a smaller number of gold-standard labels, sampled with controlled probabilities, and stays valid even when the model labels are arbitrarily biased [22]. Prediction-powered inference gives provably valid confidence intervals for means, quantiles and linear and logistic regression coefficients when a dataset is supplemented with model predictions, with no assumptions about the model, and more accurate predictions give narrower intervals [23]. Confidence-Driven Inference (2025) uses the model's labels and confidence to choose which items people label, and reached conclusions valid and no less accurate than human labels alone with over 25% fewer human annotations in each of three settings: politeness, stance and media bias [24].

Conclusions. Baumann and colleagues (2025 preprint) replicated 37 annotation tasks from 21 published social science studies, with 13 million labels from 18 models across 2,361 hypotheses [21]. Model labels led to incorrect conclusions on about 31% of hypotheses with state-of-the-art models and on half with smaller ones, and "with just a handful of prompt paraphrases, virtually anything can be presented as statistically significant" [21]. Human annotations gave crucial protection against false positives, regression correction restored valid inference at a trade between Type I and Type II errors, and the risk fell as effect sizes rose [21].

Replacing annotators. The alternative annotator test (2025) compares a model with at least three human annotators on 50 to 100 examples and judges the model to hold the advantage when its winning rate is 0.5 or more [25]. Across ten datasets, six models and four prompting techniques, commercial models could sometimes replace human annotators [25].

Training data. Pangakis and Wolken (2024) fine-tuned classifiers on model labels and on human labels for 14 tasks from social science articles, with data from password-protected archives that are less prone to contamination [33]. The median F1 gap was 0.039, and supervised models trained on model labels differed from few-shot model labeling by a median F1 of 0.006 [33]. Two set-ups trained on model labels reached 0.746 median recall, significantly better than any model trained on human labels, while an encoder classifier trained on human labels had median precision 0.214 higher than the same classifier trained on model labels [33]. The authors suggest model labels suit tasks that put recall before precision, and note that downstream classifiers do not remove bias or poor performance in the model's labels [33].

Preference labels. Lee and colleagues (2024) measured how often a model's preference labels matched human preferences [32]. The best prompts reached 78.0% on summarization, 69.1% on helpfulness and 72.1% on harmlessness, and self-consistency "strictly degrades" alignment [32]. Against an earlier estimate of 73% to 77% agreement between people on a human preference dataset, the authors judge 78% good in absolute terms [32]. Training on the model's preferences performed comparably to training on human preferences for summarization and helpful and harmless dialogue [32].

9. Human review of model suggestions, and model use in crowd work

Schroeder, Roy and Kabbara (2025) ran a pre-registered experiment with 410 annotators and over 7,000 annotations, three assistance conditions against controls, two models and two datasets of subjective tasks [30]. Suggestions did not make annotators faster but raised their self-reported confidence, and annotators "strongly took the LLM suggestions", changing the label distribution [30]:

Measure	No assistance	With suggestions
Crowd labels (decision of 3) matching the model's suggestion	40%	81% to 87%, by presentation
Full consensus of five annotators	8%	38% (pre-highlighted label)
Model's average weighted F1 against the labels	0.47	0.79, about 0.32 higher

Source: [30], sections 5.3 and 5.4.

When labels made with model help are used to evaluate that model, its reported performance rises [30].

Gu and colleagues (2026) studied event annotation: removing irrelevant documents, merging documents about the same event and annotating the events [31]. Fully automated model annotation beat TF-IDF-based methods and Event Set Curation but was still not reliable against human experts [31]. Model help reduced the time and mental effort of variable annotation [31].

Veselovsky, Horta Ribeiro and West (2023) reran an abstract summarization task on a paid crowd platform and, from keystroke detection and a classifier for synthetic text, estimated that 33% to 46% of crowd workers used a language model [29]. That task was summarization, and the authors call generalization to other tasks unclear [29].

Implications

For teams that buy or build labeled data. Validate on the task itself, since results on one task do not carry to the next, even within one paper [5, 7, 11]. Compare model labels with human labels on the same items and guidelines, on chance-corrected agreement and on accuracy against reviewed labels [6, 7, 25]. Report precision and recall by label, since a model can score well overall and fail one class [5, 7, 9].

For expert and subjective work. Expect gaps on expert and long-text tasks, and mixed results on span tasks [12, 13, 14]. On subjective labels, check agreement by annotator group as well as overall [17, 18], and treat disagreement as information about the item [19].

For anyone who estimates with model labels. Keep a human-labeled sample drawn with known probabilities, and use a correction such as design-based supervised learning, prediction-powered inference or Confidence-Driven Inference [22, 23, 24]. Report results across several prompt wordings [21].

For reproducibility and review. Record the model, its version, the prompt, the settings, the date and repeated runs [3, 8, 20]. Keep an unassisted reference set, because suggestions shift the labels people give and raise the measured score of the model that made them [30]. In one summarization task, an estimated 33% to 46% of crowd workers used a language model [29], so a crowd baseline is worth checking for model output.

For agreement figures. Report the protocol behind every number: the scale, the items kept and the handling of abstentions and invalid outputs [27]. High agreement with people can hide shared errors, and low agreement can hide better labels [26].

Rillor's data labeling and annotation service delivers training data labeled to your schema, then reviewed, with accuracy measured on a reviewed sample and reported with the delivery.

Limits of this analysis

- Most comparisons use English social media posts, news and social science tasks [3, 5, 6, 7, 9]. One study covers Mandarin [15], and expert work covers finance, law, biomedicine, ethnography and education messages [12, 14, 26].
- Each figure holds for the model version and date its study tested, from 2021 to 2026, and a hosted model can change under the same name [20].
- The studies use different measures, so no figure is pooled and no average agreement between models and people is given.
- Several sources are preprints or a working paper [4, 5, 7, 8, 14, 15, 21, 26, 27, 29].
- Human baselines are often under-reported [28]. Kasner's expert γ of 0.31 [13] and Ziems' inter-human κ of 0.51 [6] come from earlier studies as cited, and the 73% to 77% human preference agreement is an earlier estimate as Lee and colleagues report it [32].
- Gu and colleagues [31] and the survey [34] give no figures used here, and Törnberg reports results in figures only [4].
- Gilardi's arXiv abstract page still shows the first version's figures. This report uses version 2, the published one [3]. Kristensen-McLachlan and colleagues are cited from their arXiv version [9].
- Costs are study figures for their own years [1, 2, 3, 13].

Sources

1. Want To Reduce Labeling Cost? GPT-3 Can Help. Shuohang Wang, Yang Liu, Yichong Xu and others, Findings of EMNLP 2021, arXiv 2108.13487, version 1, 30 August 2021, viewed 8 November 2026. https://arxiv.org/abs/2108.13487

2. Is GPT-3 a Good Data Annotator? Bosheng Ding, Chengwei Qin, Linlin Liu and others, Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers), pp. 11173-11195, July 2023, viewed 8 November 2026. https://aclanthology.org/2023.acl-long.626/

3. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Fabrizio Gilardi, Meysam Alizadeh and Maël Kubli, Proceedings of the National Academy of Sciences 120(30), e2305016120, 2023, arXiv 2303.15056, version 2, 19 July 2023, viewed 8 November 2026. https://arxiv.org/abs/2303.15056

4. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. Petter Törnberg, working paper, arXiv 2304.06588, version 1, 13 April 2023, viewed 8 November 2026. https://arxiv.org/abs/2304.06588

5. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq and others, preprint, arXiv 2304.10145, version 2, 22 April 2023, viewed 8 November 2026. https://arxiv.org/abs/2304.10145

6. Can Large Language Models Transform Computational Social Science? Caleb Ziems, William Held, Omar Shaikh and others, Computational Linguistics 50(1), pp. 237-291, March 2024, viewed 8 November 2026. https://aclanthology.org/2024.cl-1.8/

7. Automated Annotation with Generative AI Requires Validation. Nicholas Pangakis, Samuel Wolken and Neil Fasching, preprint, arXiv 2306.00176, version 1, 31 May 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.00176

8. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. Michael V. Reiss, preprint, arXiv 2304.11085, version 1, 17 April 2023, viewed 8 November 2026. https://arxiv.org/abs/2304.11085

9. Are Chatbots Reliable Text Annotators? Sometimes. Ross Deans Kristensen-McLachlan, Miceal Canavan, Márton Kardos and others, PNAS Nexus 4(4), 2025, arXiv 2311.05769, version 2, 25 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2311.05769

10. Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. Meysam Alizadeh, Maël Kubli, Zeynab Samei and others, Journal of Computational Social Science 8(1), 2025, arXiv 2307.02179, version 2, 29 May 2024, viewed 8 November 2026. https://arxiv.org/abs/2307.02179

11. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. Xingwei He, Zhenghao Lin, Yeyun Gong and others, Proceedings of NAACL 2024 (Volume 6: Industry Track), pp. 165-190, June 2024, viewed 8 November 2026. https://aclanthology.org/2024.naacl-industry.15/

12. Evaluating Large Language Models as Expert Annotators. Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen and Hsin-Hsi Chen, COLM 2025, arXiv 2508.07827, version 1, 11 August 2025, viewed 8 November 2026. https://arxiv.org/abs/2508.07827

13. LLMs as Span Annotators: A Comparative Study of LLMs and Humans. Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová and others, MME workshop at EACL 2026, arXiv 2504.08697, version 4, 2 February 2026, viewed 8 November 2026. https://arxiv.org/abs/2504.08697

14. Large language models struggle with ethnographic text annotation. Leonardo S. Goodall, Dor Shilton, Daniel A. Mullins and Harvey Whitehouse, preprint, arXiv 2601.12099, version 1, 17 January 2026, viewed 8 November 2026. https://arxiv.org/abs/2601.12099

15. LLMs for automatic annotation of Mandarin narrative transcripts. Qingwen Zhao, Hongao Zhu, Yunqi He and others, preprint, arXiv 2605.17205, version 1, 17 May 2026, viewed 8 November 2026. https://arxiv.org/abs/2605.17205

16. GPT is Not an Annotator: The Necessity of Human Annotation in Fairness Benchmark Construction. Virginia K. Felkner, Jennifer A. Thompson and Jonathan May, Proceedings of the 62nd Annual Meeting of the ACL (Volume 1: Long Papers), pp. 14104-14115, August 2024, viewed 8 November 2026. https://aclanthology.org/2024.acl-long.760/

17. Sociodemographic Prompting is Not Yet an Effective Approach for Simulating Subjective Judgments with LLMs. Huaman Sun, Jiaxin Pei, Minje Choi and David Jurgens, Proceedings of NAACL 2025 (Volume 2: Short Papers), pp. 845-854, April 2025, arXiv 2311.09730, version 2, 17 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2311.09730

18. Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. Julian Ashwin, Aditya Chhabra and Vijayendra Rao, Sociological Methods & Research 55(3), pp. 795-839, 2025, arXiv 2309.17147, version 2, 5 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2309.17147

19. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Barbara Plank, Proceedings of EMNLP 2022, pp. 10671-10682, December 2022, viewed 8 November 2026. https://aclanthology.org/2022.emnlp-main.731/

20. How is ChatGPT's behavior changing over time? Lingjiao Chen, Matei Zaharia and James Zou, Harvard Data Science Review 6(2), 2024, arXiv 2307.09009, version 3, 31 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2307.09009

21. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation. Joachim Baumann, Paul Röttger, Aleksandra Urman and others, preprint, arXiv 2509.08825, version 2, 6 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2509.08825

22. Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. Naoki Egami, Musashi Hinck, Brandon M. Stewart and Hanying Wei, NeurIPS 2023, arXiv 2306.04746, version 3, 14 January 2024, viewed 8 November 2026. https://arxiv.org/abs/2306.04746

23. Prediction-Powered Inference. Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang and others, Science 382(6671), 2023, arXiv 2301.09633, version 4, 9 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2301.09633

24. Can Unconfident LLM Annotations Be Used for Confident Conclusions? Kristina Gligorić, Tijana Zrnic, Cinoo Lee and others, NAACL 2025, arXiv 2408.15204, version 2, 8 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2408.15204

25. The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. Nitay Calderon, Roi Reichart and Rotem Dror, Proceedings of the 63rd Annual Meeting of the ACL (Volume 1: Long Papers), pp. 16051-16081, July 2025, viewed 8 November 2026. https://aclanthology.org/2025.acl-long.782/

26. Agreement Is Not Quality: Blind Expert Verification of Human and LLM Qualitative Coding When Human Consensus Is Not Ground Truth. Alex Liu, Lief Esbenshade, Michael Xiao and others, preprint, arXiv 2607.28890, version 1, 30 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.28890

27. Agreement Metrics for LLM-as-Judge Evaluation: What to Report and Why. Delip Rao and Chris Callison-Burch, preprint, arXiv 2606.00093, version 2, 31 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2606.00093

28. Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025. Maria Kunilovskaya, Gagan Bhatia, Lisa Sophie Albertelli and others, EMNLP 2026 Main, arXiv 2606.02255, version 3, 2 September 2026, viewed 8 November 2026. https://arxiv.org/abs/2606.02255

29. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. Veniamin Veselovsky, Manoel Horta Ribeiro and Robert West, preprint, arXiv 2306.07899, version 1, 13 June 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.07899

30. Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks. Hope Schroeder, Deb Roy and Jad Kabbara, Findings of the ACL: ACL 2025, pp. 25771-25795, July 2025, viewed 8 November 2026. https://aclanthology.org/2025.findings-acl.1323/

31. Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators. Feng Gu, Zongxia Li, Carlos Rafael Colon and others, Findings of ACL 2026, arXiv 2503.06778, version 3, 28 April 2026, viewed 8 November 2026. https://arxiv.org/abs/2503.06778

32. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. Harrison Lee, Samrat Phatale, Hassan Mansoor and others, Proceedings of the 41st International Conference on Machine Learning, PMLR 235:26874-26901, 2024, arXiv 2309.00267, version 3, 3 September 2024, viewed 8 November 2026. https://arxiv.org/abs/2309.00267

33. Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. Nicholas Pangakis and Samuel Wolken, Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science, arXiv 2406.17633, version 1, 25 June 2024, viewed 8 November 2026. https://arxiv.org/abs/2406.17633

34. Large Language Models for Data Annotation and Synthesis: A Survey. Zhen Tan, Dawei Li, Song Wang and others, EMNLP 2024 Main, arXiv 2402.13446, version 3, 2 December 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.13446

About this report

Published by Rillor on 13 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Language models as annotators: agreement with people in published studies. 13 November 2026. https://rillor.com/insights/language-models-as-annotators

Work with Rillor

Send a sample of the records, the labels your schema sets and the accuracy you need. Contact us at rillor.com/contact?type=labeling or contact@rillor.com. More on labeling and annotation: rillor.com/data#labeling