

Method	F ₁
Oracle	74.12
Random	65.12
Exp.	69.78
MV	66.80
Exp.+MV	68.29

Table 5: The overall span-level proportional F₁ scores of all methods with the feedback simulator.

A. Feedback Simulator

The performance of crowd workers may fluctuate on different kinds of annotation tasks. To validate the **Exp.+MV** worker selection method under more stable settings, we exclude the datasets in the worker selection process and directly generate the numerical feedback when workers give annotations. Specifically, for each worker w , we precalculate the average F₁ score of all annotations by w on the original dataset using expert and MV evaluation respectively, denoted as $\bar{\varphi}_w^{Exp.}$ and $\bar{\varphi}_w^{MV}$. At time step t , for each $s_i \in S_t$, we assign K tasks to K different workers in W_t , then use a random value on $[0, 1]$ as the workers' agreement κ . If $\kappa > \tau$, we generate feedback for the K workers from $Bernoulli(\bar{\varphi}_w^{MV})$ independently. Otherwise, the feedback is generated from $Bernoulli(\bar{\varphi}_w^{Exp.})$. We set the kappa threshold value τ to 0.4 in **Exp.+MV**. The results of this experiment are shown in Table 5. **Exp.+MV** saves 59.88% of expert usage under these settings.

B. Regret Analysis

We provide a brief regret analysis of the worker selection framework assuming that we use the ϵ -greedy algorithm and that each worker's reward follows a Bernoulli distribution.

The main proof follows the proof of Theorem 1 in (Garcelon et al., 2022). The key contribution here is that we need to specify that the evaluation signal (generated by majority voting) is a generalized linear model of workers' true reward signal (generated by expert/oracle). To this end, we utilize the following form of the Chernoff bound which applies for any random variables with bounded support.

Lemma 1 (Chernoff Bound (Motwani and Raghu-
van, 1995)) *Let X_1, X_2, \dots, X_N be independent random variables such that $x_l \leq X_i \leq x_h$ for all $i \in \{1, 2, \dots, N\}$. Let $X = \sum_{i=1}^N X_i$ and $\mu = \mathbb{E}(X)$. Given any $\delta > 0$, we have the following result:*

$$P(X \leq (1 - \delta)\mu) \leq e^{-\frac{\delta^2 \mu^2}{N(x_h - x_l)^2}}. \quad (9)$$

For the purpose of our discussion, let $X_i \in \{0, 1\}$ be a binary random variable, where $X_i = 0$ denotes that worker i provides an incorrect solution, and $X_i = 1$ denotes that worker i generates a correct solution. Define $X = \sum_{i \in \mathcal{N}} X_i$.

We aim to approximate P_{MV} , which is the probability that the majority of the N workers provide the correct estimate.

We apply the Chernoff Bound in Lemma 1 to P_{MV} . We can compute

$$\mathbb{E}(X) = \bar{p} = \frac{\sum_{i=1}^N p_i}{N}. \quad (10)$$

Based on (9), we let $\mu = \mathbb{E}(X)$, $\delta = \frac{N(\bar{p} - \frac{1}{2})}{\frac{N}{2} + N(\bar{p} - \frac{1}{2})}$, $x_l = 0$, $x_h = 1$, and get the following result:

$$P_{MV} = P\left(X \geq \frac{N}{2}\right) = 1 - P\left(X \leq \frac{N}{2}\right) \geq 1 - e^{-\frac{\delta^2 \mu^2}{N}} \quad (11)$$

$$= 1 - e^{-\frac{N^2(\bar{p} - \frac{1}{2})^2}{[\frac{N}{2} + N(\bar{p} - \frac{1}{2})]^2} \cdot \frac{[\frac{N}{2} + N(\bar{p} - \frac{1}{2})]^2}{N}} \quad (12)$$

$$= 1 - e^{-\frac{N^2(\bar{p} - \frac{1}{2})^2}{N}} \quad (13)$$

$$= 1 - e^{-N\left(\frac{\sum_{i=1}^N p_i}{N} - \frac{1}{2}\right)^2}. \quad (14)$$

Through approximating P_{MV} by its lower bound in (14), we can see that the evaluation signal (represented by P_{MV}) is an increasing function in each worker's capability p_i and twice-differentiable. That is, P_{MV} is a generalized linear function, which satisfies Assumption 3 in (Garcelon et al., 2022). Therefore, one can follow the proof of Theorem 1 in (Garcelon et al., 2022) that the ϵ -greedy algorithm yields a sub-linear regret with order $\tilde{O}(T^{2/3})$.

C. Case Study of Annotation Errors

Based on our statistical analysis of the Chinese OEI dataset, we find that 74.80% of annotations have different types of errors. We use modifications (SES) of augmentations to simulate varying annotation errors made by human annotators. In this section, we provide a detailed characterization of human-made errors observed on annotated data with real cases to better motivate our modifications.

Shifting Some error spans are as long as expert ones, but of which the positions are wrong. *Shifting* simulates this type of error. We can see in Figure 6, both the expert span and the crowd span are three words long and of negative polarity. The difference is that the crowd span is shifted to the left by 2 words compared with the expert span. This type of error can be generated with *Shifting*.

Expert	如果你感到有些沮丧或失落，你不妨试试运动。 If you feel slightly depressed or lost, you could try sports.
Crowd Worker	如果你感到有些沮丧或失落，你不妨试试运动。 If you feel slightly depressed or lost, you could try sports.

Figure 6: A case in which the crowd worker annotates a span with correct length and polarity but incorrect position.

Expanding *Expanding* is used to generate longer (than expert span) error spans. It might be intuitive that annotators barely make errors such as expanding to a very long span. However, in the case shown in Figure 7, the expert annotates five short spans separated by commas. While the crowd worker uses a very long span that covers the whole sentence, which is obviously not accurate. To simulate such a human-made error, we can expand an expert span to cover the sentence. Statistically, 4.03% of annotation errors are very long spans with more than 15 Chinese characters. So we do not set an upper bound of span length in *Expanding*.

Expert	良好的身体素质是高效的保障，是成功的保障，是高水准生活的保障，是为社会多做贡献的保障，是生命的保障！ Good physical fitness is the guarantee of efficiency, of success, of a high standard of living, of contributing more to society, of life!
Crowd Worker	良好的身体素质是高效的保障，是成功的保障，是高水准生活的保障，是为社会多做贡献的保障，是生命的保障！ Good physical fitness is the guarantee of efficiency, of success, of a high standard of living, of contributing more to society, of life!

Figure 7: A case in which the crowd worker uses a very long span that covers the whole sentence.

Shrinking *Shrinking* is useful since crowd workers often ignore some words when annotating. As shown in Figure 8, the crowd worker failed to find all words expressing positive opinions.

Expert	我昨天在家里过得高效而又充实。 I had a productive and fruitful day yesterday at home.
Crowd Worker	我昨天在家里过得高效而又充实。 I had a productive and fruitful day yesterday at home.

Figure 8: A case in which the crowd worker does not annotate all words with polarity.

Sometimes crowd workers ignore a whole span in expert annotations. This is why we set the lower bound of span length to zero in *Shrinking*, which means we can shrink a span into no span.

These three types of modifications can be combined to simulate more complex annotation errors.

Worker ID	Ori. F ₁	Rnd. Gen. F ₁	SES Only F ₁	SES +Alg.2 F ₁	Worker ID	Ori. F ₁	Rnd. Gen. F ₁	SES Only F ₁	SES +Alg.2 F ₁
25	62.90	60.07	69.59	62.89	37	37.15	96.10	26.79	37.16
32	60.87	41.37	68.79	60.87	13	36.19	31.62	25.14	36.20
42	53.88	4.37	66.57	53.88	20	36.11	71.44	25.02	36.12
5	52.07	50.74	60.76	52.06	64	35.97	65.66	25.39	35.97
55	50.70	30.24	61.13	50.70	63	35.22	75.40	24.73	35.22
2	50.53	91.99	60.92	50.53	6	35.15	65.74	25.00	35.16
52	50.08	41.93	60.91	50.08	10	34.63	51.28	25.08	34.64
17	49.82	43.73	35.82	49.82	66	33.75	60.98	24.99	33.75
57	49.25	13.17	35.59	49.25	53	32.90	27.51	24.78	32.89
11	49.04	53.71	35.19	49.03	4	32.72	8.40	24.77	32.72
26	48.89	5.17	35.59	48.82	21	32.19	73.47	24.78	32.19
36	48.71	15.53	35.27	48.70	62	32.16	48.71	24.89	32.16
46	48.67	44.84	35.19	48.67	1	32.10	34.42	24.96	32.10
29	48.60	95.39	35.21	48.60	41	31.94	77.55	24.88	31.93
35	47.07	23.64	35.34	47.07	51	31.78	68.07	24.85	31.78
49	46.80	60.30	35.27	46.80	31	31.61	29.44	24.59	31.61
54	45.63	18.74	34.45	45.64	8	31.05	28.55	24.76	31.05
14	45.13	60.99	34.54	45.13	67	30.91	95.51	24.22	30.91
43	44.93	34.91	33.72	44.93	58	30.70	21.64	23.96	30.70
7	44.37	23.89	33.50	44.37	65	30.61	4.51	24.17	30.60
59	44.36	72.37	33.61	44.37	38	30.47	4.82	24.11	30.47
23	43.38	4.85	33.58	43.38	28	29.86	2.63	24.00	29.86
56	43.37	41.96	33.31	43.37	45	29.38	36.13	24.15	29.38
0	41.60	66.81	28.19	41.61	30	28.70	61.16	21.88	28.71
18	41.40	31.53	28.56	41.40	15	25.73	38.92	21.40	25.73
16	41.31	57.13	28.03	41.31	19	24.69	4.39	21.31	24.70
22	41.05	85.83	28.21	41.06	44	23.42	7.15	21.08	23.42
47	40.78	82.33	27.91	40.78	9	22.88	96.22	21.22	22.89
61	40.22	12.20	28.44	40.22	33	22.36	29.89	19.50	22.36
40	40.01	84.98	28.38	40.02	39	20.69	57.73	19.26	20.69
50	39.35	56.04	28.64	39.35	69	20.39	63.02	19.26	20.40
27	38.77	34.07	27.87	38.77	3	17.12	28.70	18.66	17.13
48	38.35	23.77	27.57	38.35	24	16.96	42.73	18.68	16.98
34	38.29	5.69	28.08	38.30	68	14.53	13.63	7.69	14.53
12	37.96	85.14	27.44	37.96	60	13.66	22.69	8.15	13.66

Table 6: Comparisons between different data augmentation methods on the span-level exact F₁ score of every crowd worker. **Ori.** stands for the original score in real datasets before any augmentation. **Rnd. Gen.** is a naive augmentation method with random generated annotations. **SES Only** indicates the *shifting*, *shrinking*, and *expanding* method we proposed. **SES + Alg.2** means SES with Algorithm 2 which is our final method.

Method	Token-level			Span-level Exact			Span-level Prop.		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
Oracle	62.88	68.62	64.80	54.48	51.97	53.07	72.79	64.07	68.15
Random	58.49	57.30	57.42	43.99	35.50	39.18	69.01	52.36	59.55
ϵ -G (Exp.)	61.91	64.58	62.61	51.72	46.37	48.76	72.28	60.25	65.72
ϵ -G (MV)	60.87	63.52	61.55	48.72	44.66	46.37	70.15	58.94	64.05
ϵ -G (Exp.+MV)	61.76	64.46	62.47	49.14	45.35	46.96	71.21	59.92	65.08
TS (Exp.)	62.66	64.91	63.20	49.76	42.34	45.69	72.15	60.20	65.63
TS (MV)	59.82	61.90	60.25	44.81	40.71	42.36	67.72	56.05	61.34
TS (Exp.+MV)	61.66	64.03	62.23	47.20	42.36	44.49	70.66	59.07	64.35
CUCB (Exp.)	63.02	63.75	62.93	52.24	45.51	48.56	73.05	59.53	65.60
CUCB (MV)	61.94	62.09	61.55	49.57	44.39	46.66	71.22	57.59	63.68
CUCB (Exp.+MV)	62.83	63.62	62.75	51.31	45.60	48.16	72.48	59.33	65.25

Table 7: Detailed P, R, and F₁ scores of all methods on the Chinese OEI dataset.