A Related Work

A.1 In-context learning

In-context learning (ICL) has become a new paradigm for natural language processing (NLP), where large language models make predictions only based on contexts augmented with a few examples (Dong et al., 2023; Xie et al., 2022; Shin et al., 2022; Zhang et al., 2023; Bai et al., 2023). A series of works attempts to revise, enhance, and understand ICL, which include but are not limited to prompt tuning (Kim et al., 2022; Wang et al., 2022a; Mishra et al., 2022), analyzing intrinsic mechanism (Bansal et al., 2022; Chan et al., 2022; Li et al., 2023; Garg et al., 2022), evaluations (Srivastava et al., 2023; Wang et al., 2022b), applications in multiple domains (Chen et al., 2022; Lee et al., 2022; Cho et al., 2023), and etc. Different from them, this paper studies selective annotations for ICL, which can effectively reduce the annotation cost in ICL. Furthermore, compared with recent work (Su et al., 2023), as discussed in the main paper, this work is superior in many aspects, such as the end-to-end manner, mitigation of the trade-off between diversity and representativeness, theoretical guarantees, and better empirical performance.

A.2 Coreset selection

Coreset selection focuses on selecting a small but highly informative subset from a large dataset for follow-up tasks, which can significantly reduce the data storage cost and training consumption (Huang et al., 2018; 2023; Feldman & Zhang, 2020; Sorscher et al., 2022). Most of the works on coreset selection target the scenes of supervised learning and classification (Sener & Savarese, 2018; Toneva et al., 2019; He et al., 2023). Only a few works extend coreset selection into unsupervised cases (Sorscher et al., 2022; Su et al., 2023). This paper studies unsupervised data selection for annotations in ICL, which reduces the annotation expenses of prompts and helps large language models become better few-shot learners. Also, it enjoys theoretical support. Therefore, this work is different from previous efforts and contributes to the research community.

A.3 Data distillation

Data distillation (Wang et al., 2018; Zhao et al., 2021; Shin et al., 2023; Cui et al., 2023; Du et al., 2023; Loo et al., 2023) is an alternative approach for dataset compression and curation, which is inspired by knowledge distillation. Different from coreset selection, this series of works target *synthesizing* a small but informative dataset as an alternative to the original dataset. However, data distillation is criticized for only synthesizing a small number of data points due to computational source limitations (Xia et al., 2023; Yang et al., 2023). The performances of data distillation and data selection are therefore not compared directly. Besides, it is under-explored about how to perform data distillation in an unsupervised manner on natural language processing tasks. Based on this analysis, the data distillation strategy is not involved in empirical evaluations.

B Proofs

B.1 Preliminary theoretical results

We first present some preliminary theoretical results and their corresponding proofs for the subsequential proofs of Proposition 1 and Theorem 2.

B.1.1 Lemma 1

Lemma 1. *Given a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{P})$, if the influence function meets Condition 1, then for $\forall \mathcal{S}_i, \mathcal{S}_j \subseteq \mathbf{V}$:*

$$f_{\mathcal{G}}(\mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_j) \leq \sum_{\mathbf{v} \in \mathcal{S}_i - \mathcal{S}_j} \psi_{\mathbf{v}}(\mathcal{S}_j) - \sum_{\mathbf{v} \in \mathcal{S}_j - \mathcal{S}_i} \psi_{\mathbf{v}}(\mathcal{S}_i \cup \mathcal{S}_j - \mathbf{v}), \tag{5}$$

where $\psi_{\mathbf{v}}(\mathcal{S}_j) := f_{\mathcal{G}}(\mathcal{S}_j \cup \mathbf{v}) - f_{\mathcal{G}}(\mathcal{S}_j)$.

Proof. The proof is inspired by (Rolnick & Weed, 2014). We first let

$$\mathcal{S}_i - \mathcal{S}_j = \{\mathbf{a}_1, ..., \mathbf{a}_r\} \tag{6}$$

and

$$\mathcal{S}_j - \mathcal{S}_i = \{\mathbf{b}_1, ..., \mathbf{b}_q\}, \tag{7}$$

where $r \in \mathbb{N}_+$ and $q \in \mathbb{N}_+$. According to Eq. (6), for the subsets $\mathcal{S}_i$ and $\mathcal{S}_j$, we have

$$\mathcal{S}_j \cup \mathcal{S}_i = \mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_r\}. \tag{8}$$

Afterward, we obtain

$$f_{\mathcal{G}}(\mathcal{S}_j \cup \mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_j) = f_{\mathcal{G}}(\mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_r\}) - f_{\mathcal{G}}(\mathcal{S}_j). \tag{9}$$

At a high level, Eq. (9) is to calculate the influence improvement of $\mathcal{S}_j$ after adding data points $\{\mathbf{a}_1, ..., \mathbf{a}_r\}$ into $\mathcal{S}_j$. As the influence improvement of adding one sequence of data points is equal to the sum of the influence improvement at each step, we have,

$$\begin{aligned} & f_{\mathcal{G}}(\mathcal{S}_j \cup \mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_j) \\ &= f_{\mathcal{G}}(\mathcal{S}_j \cup \mathbf{a}_1) - f_{\mathcal{G}}(\mathcal{S}_j) + \sum_{k=2}^{r} [f_{\mathcal{G}}(\mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_k\}) - f_{\mathcal{G}}(\mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_{k-1}\})] \\ &= \psi_{\mathbf{a}_1}(\mathcal{S}_j) + \sum_{k=2}^{r} \psi_{\mathbf{a}_k}(\mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_{k-1}\}). \end{aligned} \tag{10}$$

Under Condition 1, as $\mathcal{S}_j \subset \mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_{k-1}\}$, we have

$$\begin{aligned} f_{\mathcal{G}}(\mathcal{S}_j \cup \mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_j) &= \psi_{\mathbf{a}_1}(\mathcal{S}_j) + \sum_{k=2}^{r} \psi_{\mathbf{a}_k}(\mathcal{S}_j \cup \{\mathbf{a}_1, ..., \mathbf{a}_{k-1}\}) \\ &\leq \sum_{k=1}^{r} \psi_{\mathbf{a}_k}(\mathcal{S}_j) = \sum_{\mathbf{a} \in \mathcal{S}_i - \mathcal{S}_j} \psi_{\mathbf{a}}(\mathcal{S}_j). \end{aligned} \tag{11}$$

Similarly,

$$\begin{aligned} & f_{\mathcal{G}}(\mathcal{S}_j \cup \mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_i) \\ &= \psi_{\mathbf{b}_1}(\mathcal{S}_i) + \sum_{k=2}^{q} \psi_{\mathbf{b}_k}(\mathcal{S}_i \cup \{\mathbf{b}_1, ..., \mathbf{b}_{k-1}\}) \geq \sum_{k=1}^{q} \psi_{\mathbf{b}_k}(\mathcal{S}_i \cup \mathcal{S}_j - \mathbf{b}_k) = \sum_{\mathbf{b} \in \mathcal{S}_j - \mathcal{S}_i} \psi_{\mathbf{b}}(\mathcal{S}_i). \end{aligned} \tag{12}$$

By subtracting (12) from (10), we have

$$f_{\mathcal{G}}(\mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_j) \leq \sum_{\mathbf{v} \in \mathcal{S}_i - \mathcal{S}_j} \psi_{\mathbf{v}}(\mathcal{S}_j) - \sum_{\mathbf{v} \in \mathcal{S}_j - \mathcal{S}_i} \psi_{\mathbf{v}}(\mathcal{S}_i \cup \mathcal{S}_j - \mathbf{v}). \tag{13}$$

□

B.1.2 LEMMA 2

Lemma 2. *Given a graph* $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{P})$*, for any subset* $\mathcal{S} \subset \mathbf{V}$ *and any* $\mathbf{v} \in \mathbf{V}$*, the influence function* $f_{\mathcal{G}}$ *satisfies*

$$\psi_{\mathbf{v}}(\mathcal{S}) = f_{\mathcal{G}}(\mathcal{S} \cup \mathbf{v}) - f_{\mathcal{G}}(\mathcal{S}) \geq 0 \tag{14}$$

Proof. We consider two cases to finish the proof.

Case 1 ($\mathbf{v} \in \mathbf{V} \wedge \mathbf{v} \notin \mathcal{S}$). In this case, the influence improvement is at least 1 since $\mathbf{v}$ itself has been included, i.e.,

$$\psi_{\mathbf{v}}(\mathcal{S}) = f_{\mathcal{G}}(\mathcal{S} \cup \mathbf{v}) - f_{\mathcal{G}}(\mathcal{S}) \geq 1. \tag{15}$$

Case 2 ($\mathbf{v} \in \mathbf{V} \wedge \mathbf{v} \in \mathcal{S}$). In this case, the influence improvement is 0 since $\mathbf{v}$ has already been included in $\mathcal{S}$, i.e.,

$$\psi_{\mathbf{v}}(\mathcal{S}) = f_{\mathcal{G}}(\mathcal{S} \cup \mathbf{v}) - f_{\mathcal{G}}(\mathcal{S}) = 0. \tag{16}$$

Combining the above two cases, we conclude that, for $\forall \mathbf{v} \in \mathbf{V}$, the influence function $f_{\mathcal{G}}$ satisfies

$$f_{\mathcal{G}}(\mathcal{S} \cup \mathbf{v}) - f_{\mathcal{G}}(\mathcal{S}) \geq 0. \tag{17}$$

□

B.2 Proof of Proposition 1

Proof. Given a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{P})$, for $\forall \mathcal{S}_i, \mathcal{S}_j \subset \mathbf{V}$, according to Lemma 2, we have

$$\sum_{\mathbf{v} \in \mathcal{S}_i - \mathcal{S}_j} \psi_{\mathbf{v}}(\mathcal{S}_i \cup \mathcal{S}_j - \mathbf{v}) \geq 0. \tag{18}$$

Taking (18) into Lemma 1, we have

$$f_{\mathcal{G}}(\mathcal{S}_i) - f_{\mathcal{G}}(\mathcal{S}_j) \leq \sum_{\mathbf{v} \in \mathcal{S}_i - \mathcal{S}_j} \psi_{\mathbf{v}}(S_j). \tag{19}$$

We use $\mathcal{S}_m^*$ to denote the optimal solution as discussed in the main paper. At any step t in Algorithm 2, we substitute $\mathcal{S}_m^*$ (resp. $\mathcal{S}_t$) into $\mathcal{S}_i$ (resp. $\mathcal{S}_j$) in (19), we can derive

$$f_{\mathcal{G}}(\mathcal{S}_m^*) \leq f_{\mathcal{G}}(\mathcal{S}_t) + \sum_{\mathbf{v} \in \mathcal{S}_m^* - \mathcal{S}_t} \psi_{\mathbf{v}}(\mathcal{S}_t). \tag{20}$$

According to Condition 1,

$$\psi_{\mathbf{v}}(\mathcal{S}_t) \geq \psi_{\mathbf{v}}(\mathcal{S}_{t+1}) \tag{21}$$

holds. Taking both (20) and (21) into (19), we have for any t,

$$f_{\mathcal{G}}(\mathcal{S}_m^*) \leq f_{\mathcal{G}}(\mathcal{S}_t) + m\psi_{t+1}. \tag{22}$$

□

B.3 Proof of Theorem 2

Proof. Recall that

$$\psi_t = f_{\mathcal{G}}(\mathcal{S}_t) - f_{\mathcal{G}}(\mathcal{S}_{t-1}). \tag{23}$$

According to Proposition 1, we have

$$f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_t) \leq m\psi_{t+1} = m(f_{\mathcal{G}}(\mathcal{S}_{t+1}) - f_{\mathcal{G}}(\mathcal{S}_t)). \tag{24}$$

Afterwards, (24) equals to,

$$\begin{aligned} & f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_t) - (f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_{t+1})) \geq \frac{1}{m}(f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_t)) \\ \iff & f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_{t+1}) \leq \frac{m-1}{m}(f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_t)). \end{aligned} \tag{25}$$

Based on (25), we have

$$\begin{aligned} f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_{t+1}) &\leq \frac{m-1}{m}(f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_t)) \\ &\leq (\frac{m-1}{m})^2(f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_{t-1})) \\ &\leq \ldots \leq (\frac{m-1}{m})^{t+1}(f_{\mathcal{G}}(\mathcal{S}_*^m) - f_{\mathcal{G}}(\mathcal{S}_0)). \end{aligned} \tag{26}$$

Since $f_{\mathcal{G}}(\mathcal{S}_0) = f_{\mathcal{G}}(\emptyset) = 0$, we have

$$\frac{f_{\mathcal{G}}(\mathcal{S}_m^*) - f_{\mathcal{G}}(\mathcal{S}_{t+1})}{f_{\mathcal{G}}(\mathcal{S}_m^*)} \leq (\frac{m-1}{m})^{t+1}. \tag{27}$$

When Algorithm 2 terminates at step $t = m - 1$, we have,

$$f_{\mathcal{G}}(\mathcal{S}_m) \geq (1 - (1 - 1/m)^m) f_{\mathcal{G}}(\mathcal{S}_m^*). \tag{28}$$

□

C SUPPLEMENTARY EXPERIMENTAL RESULTS

C.1 SELECTED EXAMPLES

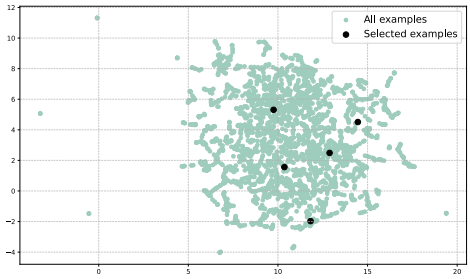
In Table 6, for illustration purposes, we provide a few examples from the selection by our method, when the annotation size is 18.

Dataset	Input
MRPC	a. Input: The two Democrats on the five-member FCC held a press conference to sway opinion against [...] Output: not equivalent a. Input: The report shows that drugs sold in Canadian pharmacies are manufactured in facilities approved by Health Canada [...] Output: equivalent c. Input: The chief merchandising officer decides what the store is going to sell [...] Output: equivalent
SST-5	a. Input: plodding, poorly written, murky and weakly acted, the picture feels as if everyone making it lost their movie mojo. Output: very negative b. Input: duvall is strong as always . Output: very positive c. Input: lohman adapts to the changes required of her , but the actress and director peter kosminsky never get the audience to break [...] Output: neutral
MNLI	a. Input: This prosperous city has many museums, including a well-endowed Musee des Beaux-Arts (Square Verdrel) [...] Output: False b. Input: Duhame, who today makes her living as a graphic designer and illustrator, calls her book [...] Output: Inconclusive c. Input: At the agency or program level, it included management's public commitment to reduce fraud and errors, as. Based on that information [...] Output: True
DBpedia	a. Input: Lars Nielsen (born 3 November 1960 in Copenhagen) is a Danish rower. Output: athlete b. Input: Calhoun County High School is a public secondary school in St. Matthews South Carolina USA. Output: educational institution c. Input: David Goldschmid (sometimes credited as Dave Goldschmid) is an American television writer and producer currently writing for the daytime drama General Hospital. Output: artist
RTE	a. Input: In sub-Saharan Africa about one in every 30 people is infected with HIV.. 30% of the people infected with HIV live in Africa.. Output: False b. Input: The drawbacks of legalization do not imply that our current version of prohibition is the optimal drug strategy; it may well [...] Output: False c. Input: For example, the fields of Western farmers feed the United States and many other parts of the world, and India's irrigation [...] Output: True
HellaSwag	a. Input: The topic is Preparing salad. An illustrated egg, the website "startcooking com" and "vegetable salad" [...] Output: is shown from above. b. Input: The topic is Pets and Animals. [header] How to treat an injured rabbit's paw [title] Identify sore hocks. [step] Pododermatitis [...] Output: Once the condition has set in, though, you'll need to take quick action to treat the injury. Leaving [...] c. Input: The topic is Playing squash. Two men stand on a racquetball court. the men Output: stretch then begin playing.

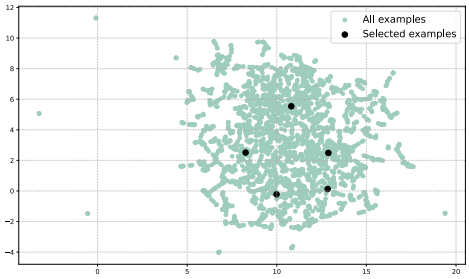
Table 6: For illustration purposes, under our method, we show randomly selected three examples from each of the six datasets in one same run (excluding the other three datasets due to their length) when the annotation budget is set to 18.

C.2 VISUALIZATION OF SELECTED EXAMPLES

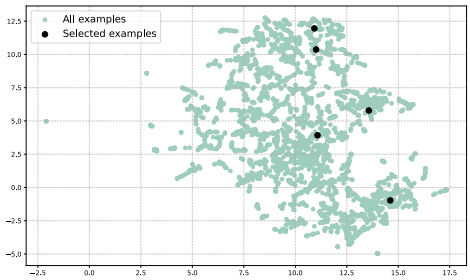
Here we provide a umap (McInnes et al., 2018) visualization of selected examples. To avoid the denseness, we choose the annotation budget as 5. The visualization can be checked in Figure 6. First, comparing subfigures (a) and (b), we can clearly see that the selection of Vote-k is much biased, and our IDEAL can identify a subset that is more favorable to be a proxy of full data. Second, comparing subfigures (c) and (d), we can see that the selected subset by Vote-k is distributed on the right of full data. By comparison, our IDEAL can select a subset that is distributed more uniformly.



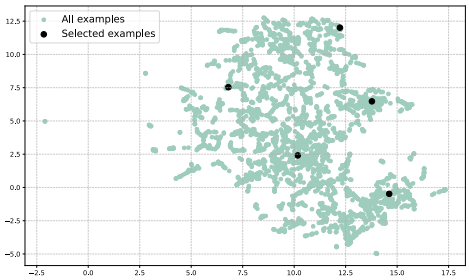
(a) SST-5, Vote-k.



(b) SST5, IDEAL.



(c) MNLI, Vote-k.



(d) MNLI, IDEAL.

Figure 6: Umap (McInnes et al., 2018) visualization to compare five selected examples from all examples using fully unsupervised methods Vote-k and IDEAL (ours). Compared with Vote-k, IDEAL could choose the examples to better represent the whole data rather than get involved in diversity and including outliers.

C.3 DETAILED EXPERIMENTAL RESULTS IN TABLE 1

Method		MRPC	SST-5	MNLI	DBpedia	RTE
100	Random	64.3/68.4/58.6	49.6/51.1/47.2	38.2/40.2/36.7	89.8/91.0/88.2	55.3/55.9/55.1
100	Vote-k	64.6/68.8/62.1	46.6/47.2/46.1	38.9/43.8/35.5	89.2/89.8/88.7	57.6/58.2/57.4
100	IDEAL	**66.4**/67.9/64.8	**51.4**/53.5/49.6	**41.0**/41.4/40.2	**90.6**/91.4/89.5	**58.9**/60.9/57.4
18	Random	57.4/68.8/39.8	42.9/46.9/39.1	37.8/39.4/35.2	85.2/87.5/83.9	57.9/58.9/57.0
18	Vote-k	61.1/67.2/52.7	41.7/45.7/37.1	39.1/43.8/32.0	89.9/94.1/87.1	58.2/58.9/57.8
18	IDEAL	**63.0**/63.7/62.5	**43.2**/45.7/39.5	**40.0**/41.8/37.1	**90.1**/90.2/89.8	**59.4**/60.9/57.8

Table 7: Mean/maximum/minimum evaluation results of all methods on classification tasks in Table 1 over three different trials. The best mean result in each case is **bolded**.

In the main paper (Table 1), we report the mean evaluation results for different methods over three random trials. Here we provided the detailed results of Table 1 with mean/maximum/minimum values. We can observe IDEAL achieves stable results compared with baselines. Moreover, the worst-case performance of IDEAL is obviously better compared with baselines in most cases.

Method		HellaSwag	MWoZ	GeoQ	Xsum
100	Random	66.7/70.3/64.1	39.9/48.4/39.9	55.3/57.8/53.1	15.3/16.4/14.8
100	Vote-k	67.9/69.9/64.0	48.3/50.8/46.9	**58.8**/60.5/57.0	17.2/17.6/16.4
100	IDEAL	**68.6**/71.9/65.2	**52.2**/55.9/49.1	58.2/60.5/54.7	**19.9**/20.2/19.5
18	Random	66.0/68.8/63.7	37.0/46.5/28.1	47.5/49.2/44.9	13.6/14.5/12.5
18	Vote-k	66.5/71.9/62.5	37.7/43.8/32.4	50.9/54.3/47.7	15.2/16.0/14.5
18	IDEAL	**67.1**/71.9/64.5	**38.5**/47.3/30.9	**52.0**/53.9/50.8	**19.6**/20.2/18.9

Table 8: Mean/maximum/minimum evaluation results of all methods on multi-choice, dialogue, and generation tasks in Table 1 over three different trials. The best mean result in each case is **bolded**.

Method	MRPC		MNLI			RTE	
	Equivalent	Not equivalent	True	Inconclusive	False	True	False
Original	2023	977	1051	965	984	1241	1249
Random	70	30	30	39	31	56	44
Vote-k	64	36	27	35	38	46	54
IDEAL	65	35	37	34	29	49	51

Table 9: The numbers of different labels in the selected examples for different methods. "Original" denotes the label statistics of the original dataset. Under the annotation budget 100, IDEAL achieves the smallest ratio between the numbers of the most frequent class and the least frequent class in 2 out of 3 cases (MNLI and RTE), implying IDEAL can indeed mitigate the label skew problem.

Method	MRPC		SST-5		RTE	
	Mean	Std	Mean	Std	Mean	Std
Random	44.2	0.02	45.4	0.02	57.3	0.02
Vote-k	52.7	0.03	38.0	0.01	58.6	0.02
IDEAL	65.5	0.01	46.6	0.01	57.5	0.02

Table 10: The average performance of different methods by permuting the order of prompts for each test instance 10 times. We conduct experiments on MRPC, SST-5, and RTE datasets and report the average results with standard deviation. We can observe the subset selected by IDEAL achieves the best performance compared with baselines in 2 out of 3 cases. IDEAL also achieves the lowest standard deviations in all evaluations, which suggests IDEAL is a more stable and robust method against the order of prompts.

C.4 Label distributions in selective annotations

Recall that the process of selective annotations is based entirely on similarities derived from sentence embeddings without labels. Therefore, we investigate whether the selected examples have label skew. Under an annotation budget of 100, we collect all selected examples in three classification tasks (MRPC, MNLI, and RTE) and show the numbers of different labels for different methods in Table 9. We also present the label statistics of the original training data. We observe that random selection shows a great variance. However, in an ideal case, it should achieve a similar distribution as the original training data. Notably, IDEAL achieves the smallest ratio between the numbers of the most frequent class and the least frequent class in 2 out of 3 cases (MNLI and RTE). This demonstrates that IDEAL can indeed balance the label distribution in the selected subset and mitigate the problem of label skew.

C.5 Prompt order in selective annotation

As pointed out by (Lu et al., 2021), the performance of in-context learning is influenced not only by the selection of prompts but also by the order in which the prompts are presented to models. Although this work focuses solely on selective annotation problems, we are interested in explor-

ing whether the selected subset can still lead to better performance when the order of prompts is permuted. Under an annotation budget of 18, we first retrieve prompts for each test instance from selected subsets achieved by different selective annotation methods. We then permute the order of prompts for each test instance 10 times, resulting in 10 different experimental trials. We show the average performance of these 10 trials and make a comparison between different selective annotation methods. We conduct experiments on MRPC, SST-5, and RTE datasets and present the results in Table 10. The results show that IDEAL outperforms baselines in 2 out of 3 cases, suggesting that our method can choose more stable and robust subsets against changed prompt orders.

D Supplementary Descriptions of Experimental Settings

D.1 Details of datasets

In this paper, to demonstrate the superiority of our method, we employ 9 datasets which can be categorized into 4 different tasks, including *classification* (MRPC (Dolan et al., 2004), SST-5 (Socher et al., 2013), MNLI (Williams et al., 2017), DBpedia (Lehmann et al., 2015), and RTE (Bentivogli et al., 2009)), *multi-choice* (HellaSwag (Zellers et al., 2019)), *dialogue* (MWoZ (Budzianowski et al., 2018)), and *generation* (GeoQuery (Zelle & Mooney, 1996) and Xsum (Narayan et al., 2018)). We list the datasets and the models used in Table 11.

	Datasets	Task	Models
Classification	MRPC (Dolan et al., 2004)	Paraphrase Detection	GPT-Neo, GPT-J, GPT-3.5-Turbo
	SST-5 (Socher et al., 2013)	Sentiment Analysis	GPT-J
	DBpedia (Lehmann et al., 2015)	Topic Classification	GPT-J
	RTE (Bentivogli et al., 2009))	Natural Language Inference	GPT-Neo, GPT-J, GPT-3.5-Turbo
	MNLI (Williams et al., 2017)	Natural Language Inference	GPT-Neo, GPT-J, GPT-3.5-Turbo
Multiple-Choice	HellaSwag (Zellers et al., 2019)	Commonsense Reasoning	GPT-J
Dialogue	MWoZ (Budzianowski et al., 2018)	Dialogue State Tracking	Text-davinci-002
Generation	GeoQuery (Zelle & Mooney, 1996)	Semantic Parsing	Text-davinci-002
	Xsum (Narayan et al., 2018)	Summarization	GPT-J

Table 11: The datasets and corresponding models used in our experiments. We use GPT-J 6B and Text-davinci-002 by default. Other large language models are explored in §4.3.4.

To help readers better understand the datasets and tasks, for each of these datasets, we also list one example including both the input and output.

D.1.1 MRPC

Input

```
Are the following two sentences 'equivalent' or 'not equivalent'?\nA
federal judge yesterday disconnected a new national \" do-not-call \"
 list , just days before it was to take effect , saying the agency
that created it lacked the authority ..\nA federal judge yesterday
struck down the national do-not-call registry slated to take effect
Oct. 1 , ruling the Federal Trade Commission had no authority to
create the list ..\nanswer:
```

Output

```
equivalent
```

D.1.2 SST-5

Input

```
How do you feel about the following sentence?\nsmug , artificial ,
ill-constructed and fatally overlong ... it never finds a consistent
tone and lacks bite , degenerating into a pious , preachy soap opera
.\nanswer:
```

Output

```
neutral
```

D.1.3 MNLI

Input

```
yeah well the Cardinals i don't know  i think the Cowboys probably
have a a better team they just at the end of the season the kind of
got messed up with Aikman getting hurt because uh Laufenberg just
couldn't never really get it together at all of course he sat along
the sidelines all season he never really got in a game never did a
whole lot. Based on that information, is the claim The Cowboys should
 have started Laufenberg all season.  \"True\", \"False\", or \"
Inconclusive\"?\nanswer:
```

Output

```
Inconclusive
```

D.1.4 DBPEDIA

Input

```
title: V\u00edctor David Loubriel; content:  V\u00edctor David
Loubriel Ort\u00edz is a Puerto Rican politician and former member of
 the Senate of Puerto Rico for the New Progressive Party (PNP).
Loubriel presented his candidacy for the Senate of Puerto Rico before
 2004. He ran for a candidate slot in the 2003 primaries obtaining
the most votes in his district (Arecibo).In the 2004 general election
 Loubriel won a seat in the 23rd Senate of Puerto Rico to represent
the district of Arecibo along with Jos\u00e9 Emilio Gonz\u00e1lez Vel
\u00e1zquez.
```

Output

```
office holder
```

D.1.5 RTE

Input

```
MEXICO CITY (Reuters) - A deadly strain of swine flu never seen
before has broken out in Mexico, killing as many as 60 people and
raising fears it is spreading across North America. The World Health
Organization said it was concerned about what it called 800 \"
influenza-like\" cases in Mexico, and also about a confirmed outbreak
 of a new strain of swine flu in the United States. It said about 60
people had died in Mexico. Mexico's government said it had confirmed
that at least 16 people had died of the swine flu in central Mexico
and that there could be another 45 fatal victims..\nquestion: 800
Mexicans have been affected by a new form of swine influenza.. True
or False?\nanswer:
```

Output

```
True
```

D.1.6 HELLASWAG

Input

```
The topic is Work World. [header] How to become a high school social
studies teacher [title] Obtain your bachelor's degree in education. [
step] All schools will require you to obtain at least your bachelor's
 degree in education. This degree will be proof that you are capable
of delivering information to students using the current educational
best practices.
```

Output

```
Make sure you've fully completed all of your course work and obtained
 your bachelor's degree before you seek certification or employment.
[substeps] Your electives should be based in social studies courses.
```

D.1.7 MULTIWOZ

Input

```
CREATE TABLE hotel(
  name text,
  ......,
  internet text CHECK (internet IN (dontcare, yes, no))
)
/*
4 example rows:
SELECT * FROM hotel LIMIT 4;
name  pricerange  type  parking book_number_of_days book_day
book_people
area  stars internet
a and b guest house moderate guest house dontcare 3 friday 5 east 4
yes
......
/*
......
-- Using valid SQLite, answer the following multi-turn conversational
questions for the tables provided above.
Example #1
[context] hotel-area: west, hotel-stars: 3, hotel-internet: yes
[system] the hobsons house is available in that area .
Q: [user] that sounds like it will work . can i book that for 3
nights
starting wednesday ?
SQL: SELECT * FROM hotel WHERE book_day = wednesday AND book_people =
 1
AND book_number_of_days = 3 AND name = hobsons house;
......
```

Output

```
hotel WHERE book_day = wednesday AND book_number_of_days = 4 AND name
 =
warkworth house;
```

D.1.8 GEOQ

Input

```
CREATE TABLE "border_info" ("state_name" text, "border" text)
/*
state_name    border
   alabama tennessee
   alabama   georgia
   alabama   florida
*/
......
-- Using valid SQLite, answer the following questions for the tables
provided above.
......
-- what is the longest river in the state with the highest point
SELECT
```

Output

```
RIVERalias0.RIVER_NAME FROM HIGHLOW AS HIGHLOWalias0, RIVER AS
RIVERalias0 WHERE HIGHLOWalias0.HIGHEST_ELEVATION = (SELECT MAX(
HIGHLOWalias1.HIGHEST_ELEVATION) FROM HIGHLOW AS HIGHLOWalias1 ) AND
RIVERalias0.TRAVERSE = HIGHLOWalias0.STATE_NAME ORDER BY RIVERalias0.
```

```
LENGTH DESC LIMIT 1;
```

D.1.9 XSUM

Input

```
For decades, large numbers of Haitians have migrated - many of them
without papers - to the Dominican Republic, to escape the poverty and
 lack of employment in their homeland.\nIn 2013, the Dominican
Republic's highest court ruled that children born there to
undocumented migrants were not automatically eligible for Dominican
nationality.
......
\nThere he strips the trees for firewood to make charcoal, to sell to
 Dominican traders for a few dollars.\nHe knows the practice damages
the fertility of the soil, but it's the only available source of
income.\n\"This is the only way we can survive,\" he says, motioning
at his family, stuck inside the world's forgotten migrant crisis.\
nYou can hear more of Will Grant's report on Heart and Soul on the
BBC World Service.
```

Output

```
Immigration has long been a divisive issue on Hispaniola, the
Caribbean island shared by Haiti and the Dominican Republic.
```

D.2 IMPLEMENTATION DETAILS

General experimental conditions. We primarily use PyTorch (Paszke et al., 2019) to implement our algorithm and baselines. For GPT-3.5-Turbo, we perform the experiments by calling the OpenAI API using a single Intel Xeon CPU. The GPT-J 6B and GPT-Neo 2.7B models are from the Huggingface transformer library (Wolf et al., 2019). We run all our experiments of GPT-J 6B and GPT-Neo 2.7B on a single NVIDIA Tesla V100 (32GB) GPU.

Details of getting unlabeled data. Since obtaining unlabeled examples in realistic scenarios is also a high-variance process, we follow the same setting as (Su et al., 2023) to simulate the realistic setting. We perform selective annotations from 3k instances that are randomly sub-sampled from training data for each task. For each experiment, we repeat the sub-sampling process three times and average the results over all trials to ensure comprehensive evaluations.

Details of the graph construction. Except for the illustration experiment in Figure 1, we construct the directed graph for all unlabeled data by connecting each vertex to its 10 nearest successors ($k = 10$). It is important to note that a larger k will lead to an increase in the computation cost. We have chosen this setting because it provides good performance while maintaining efficient computation costs. For Figure 1, we construct the graph by connecting each vertex to its 3 nearest successors in order to avoid denseness ($k = 3$).

Details of Algorithm 1. Considering the randomness of the diffusion process, when quantifying the influence of the subset, we run Algorithm 1 10 times and use the averaged influence value. Note that we also calculate the time cost in this repeated process when reporting the final results in the main paper. As shown in Figure 3, our algorithm is still more effective than Vote-k.

E LIMITATIONS

Memory cost. Although in-context learning tasks avoid the heavy parameter update process, they still require a large amount of memory to load models. For example, loading GPT-J 6B into a GPU requires about 23GB GPU memory, without considering the size of the dataset. This is a relatively high cost for individual researchers.

Time cost of Auto-IDEAL. Although Auto-IDEAL achieves even better performance than IDEAL, it has the same drawback as Vote-k. That is to say, when making automatic annotations, it incurs the

cost of making predictions for all unlabeled data. Future work may study how to maintain superior performance while reducing the automatic annotation cost of IDEAL at the same time. Compared with (Su et al., 2023), we do not evaluate NQ (Kwiatkowski et al., 2019) due to budget constraints.