Multi-Class Data Labeling through Large Language Models and Weak Supervision

Anonymous ACL submission

Abstract

Recent growing interest in the fields of artificial intelligence and machine learning has underscored the need for high-quality labeled datasets that can be used to train machine learning models and empower large language models to provide specialized knowledge. However, obtaining such a labeled dataset manually can be a time-consuming, expensive process. In this paper, we study the specific issue of labeling a multi-class dataset, and develop two general automated solutions to this issue: one using a weakly supervised NLP approach based on semantic similarity, and the other using large language models and a prompting approach called in-context learning. We run a series of experiments on an already-labeled dataset of research articles associated with multiple disciplines. Our proposed method, multi-class prompting with in-context learning, or MultiICL, leverages the multi-label nature of the dataset in its labeling. This method outperforms other methods and proves to be the best choice for multi-class data labeling.

1 Introduction

Innovations in artificial intelligence and machine learning, particularly with the rise of large language models (LLMs), are arriving at a blistering rate. Accompanying this growing interest in AI/ML technology is a need for high-quality labeled data, particularly for supervised learning applications. From medical imaging to spam detection, there are many applications of supervised learning that require these labeled datasets (Crawford et al., 2015) (Erickson et al., 2017). Many of the applications don't have existing datasets, and often require experts to assist with the labeling, a process that can be time-consuming and expensive (Masud et al., 2008).

High-quality labeled training data is also crucial for developers interested in building custom large language models to serve as "experts" for a particular subject. These would be fine-tuned foundation models like GPT-4 and LLaMA, which require a significant amount of such data to achieve optimal performance.

For example, one group of researchers saw the need for a comprehensive corpus of climate research (Mallick et al., 2023), categorized to address various climate hazards and research topics. Since these research articles can belong to multiple categories, this was a multi-label classification problem, where, given a set of research articles, the researchers had to label them into one or more categories. In order to train a classifier to perform this task, they needed a training set of already-labeled articles. Climate research is not the only field in which generating labeled datasets for machine learning applications is valuable. In medicine, researchers used advanced information retrieval techniques to identify medical misinformation on health forums (Kinsora et al., 2017). Researchers also investigated the data labeling problem for image and audio classification (Mithun et al., 2016) (Fonseca et al., 2021). While much research has prioritized human-labeled datasets, there has been growing interest in automated methods for data labeling.

To that end, in this paper, we propose MultiICL, a framework for labeling multi-class datasets using LLMs with in-context learning and a novel "multi-class prompting" approach. MultiICL performs well across many tasks, and requires only a fraction of the LLM requests that other LLM-based methods use for multi-class labeling. We compare this approach with many automated methods across 15 experiments. One such approach is the weakly supervised NLP approach with semantic similarity, devised by (Mallick et al., 2023). We also study a range of open-source and closed-source LLMs of multiple sizes, as well as evaluate the effect of various prompt engineering techniques

like in-context learning, chain-of-thought reasoning, and the multi-class prompting approach. For the first time, we also apply weak supervision, normally used with human-created labeling functions, to LLMs, treating them as noisy sources of supervision and combining them into a single weak supervision model using Snorkel. We evaluate these models on a multi-class dataset of research articles. Each article can belong to multiple fields, such as mathematics and computer science. There is significant overlap between some fields.

Our two biggest contributions are as follows. First, we evaluate the semantic similarity and weak supervision method for the first time, using a method in which class definitions are used to label documents by computing their similarity to the document. Second, as far as we know, we are the first study to examine data labeling for a multi-class dataset. This problem inspires our favored approach, which we call MultiICL.

2 Related work

The issue of obtaining high-quality, labeled data is pervasive across the machine learning community, and has motivated a large body of research that has investigated automated approaches to labeling data. In this section, we will review research about two potential solutions to this problem, weak supervision and LLMs, and discuss what gaps within this body of research our work fills.

Weak supervision is an approach to creating labeled data in which many noisy "labeling functions" are combined to create large training sets. These labeling functions (LFs), which can be as simple as checking for keywords in text, can be combined in a probabilistic model that learns where the LFs agree and disagree to produce the best labels. This paradigm, called data programming (Ratner et al., 2016), was pioneered by researchers at Stanford with their Snorkel model, and has exhibited strong performance. A case study by these researchers found that having humans create LFs instead of manually labeling the data results in a process that is 2.8x faster and has a 45.5% higher predictive performance (Ratner et al., 2017). Another study attempting to extract data from legal documents and train classifiers found similar results, with high precision, recall, and F1-scores across all models (Elwany et al., 2022).

These results are encouraging, but one limitation of weak supervision is that creating these labeling functions can often be hard, requiring creativity and domain expertise. One group of researchers developed a method they call "interactive weak supervision" (IWS). In this approach, users are actively asked about potential LFs, which are sourced using NLP methods and sampling text; an IWS model then learns from feedback on these LFs to choose the best final set of them. Experiments with real users reveal that humans can recognize high accuracy LFs and annotate these heuristics faster than individual data entries (Boecking et al., 2020).

Other researchers have addressed this problem by developing non-keyword-based labeling functions. NLP methods have proven effective: one group of researchers identify suicidal ideation from clinical notes using a rule-based NLP algorithm to identify negative examples (Cusick et al., 2021). In the medical field as well, one group developed Trove, a weakly supervised biomedical named entity recognition (NER) model to identify concepts like disease names and associated attributes from medical ontologies. Their research discusses one issue with hand-labeling data, which is that these datasets are static and expensive to change; in the context of the rapid, unexpected COVID-19 pandemic, more flexible machine learning tools are necessary (Fries et al., 2020).

As mentioned, our research adds to this body of work by evaluating another NLP method as a labeling function for weak supervision: semantic similarity. In this method (described in more detail in the Methodology section), if the similarity score between the outcome text and another text exceeds a certain threshold, then the outcome text is positively classified.

The arrival of capable large language models has resulted in numerous applications, one of which is data labeling. While LLMs can be expensive to use, training a less advanced model on training data labeled by LLMs can actually outperform the LLMs themselves, while being more cost-effective (Wang et al., 2021).

Research has generally supported the fact that LLMs are effective at labeling data. AI startup Refuel has found that LLMs generally are around 20x faster and 7x cheaper than human annotators (Team). Other work has found that using ChatGPT-4 and other LLMs to label diverse sets of data, from Twitter messages to news articles, is substantially faster and more accurate than crowd workers like those from Amazon's Mechanical Turk (MTurk) (Törnberg, 2023) (Gilardi et al., 2023). These find-

ings are not uniform, however: a study labeling tweet contents finds lower accuracy using ChatGPT, although it concludes that these LLMs have the potential to be better (Zhu et al., 2023). Other research has pointed to limitations of certain models and in multilingual tasks (Mohta et al., 2023).

Research has also highlighted the importance of prompt engineering techniques like in-context/few-shot learning and chain-of-thought reasoning. In light of recent research showing that language models can learn effectively from only a few examples (Brown, 2020), some articles have investigated how this capability can be applied to labeling data. In-context learning (ICL), where labeled examples are included in the prompt itself, have been shown to be effective; research done by Refuel has found that ICL boosts label quality by an average of 7 percentage points (Team).

Another paper outlines a novel approach to using both in-context learning and chain-of-thought (CoT) prompting, another technique where models are encouraged to use intermediate reasoning steps before arriving at an answer (Wei et al., 2022). In the system, which the researchers call AnnoLLM, LLMs are prompted to give an explanation for why a specific label is given, then that reason is used alongside the labeled example in an ICL prompt. Like other research has shown, this approach surpasses crowdsourced annotations (He et al., 2023).

Our research builds off of this larger body of research in a few ways. Our experiments examine the importance of size, contrasting smaller LLMs (Mistral, Llama 3 8B) with larger LLMs (GPT-4, Llama 3 70B) and seeing how prompting techniques like in-context learning can affect the classification at different size levels.

Most notably, we use a multi-label dataset, which allows me to examine a unique prompting approach called MultiICL, in which labeled examples and instructions to label documents into multiple categories are included in a single prompt. Using a multi-label dataset also yields a number of interesting results, in particular the insights from how predicted labels overlap with each other, which helps us understand the misclassifications better.

It is certainly possible to combine both weak supervision and LLMs, and research conducted by Snorkel AI has shown that this approach is effective, especially with smaller, noisier models. When combined with human-created labeling functions and providing a few shots of labeled examples to fine-tune the LLM, they find that the Snorkel model can produce results with a very high accuracy (Smith, 2023). Another group of researchers uses weak supervision and LLMs in a completely different way: instead of combining human-created labeling functions and a fine-tuned LLM, they give multiple prompts to the LLM concerning different aspects of the text; its answers will result in a collection of possible labels, which will be fed into the Snorkel model to produce the best label. This multiple-prompt approach improves performance on the WRENCH weak supervision benchmark (Zhang et al., 2021) by 19.5% over zero-shot prompting (where only one prompt is asked) (Smith et al., 2024).

As mentioned, we also apply weak supervision to the LLM side of the research, evaluating three small LLMs within a weak supervision model to see if performance improves. While others have used the methods of weak supervision and LLMs together, no one, as far as we know, has examined treating LLMs themselves as sources of supervision.

3 Methodology

In this section, we describe the various methods used for automated document labeling. These include weak supervision and LLMs with in-context learning, chain-of-thought reasoning, and multi-class prompting. We also discuss applying the weak supervision paradigm to LLMs.

3.1 Weak supervision

In our weak supervision with semantic similarity approach, we first define or identify a definition for the research field and then determine its similarity to the abstract of a research article using sentence embeddings. We employ pretrained sentence transformers to convert text into vector representations that encode semantic information, and then calculate the cosine similarity between the definition and abstract vectors. If the similarity score exceeds a threshold (in these experiments, 0.5), the abstract is classified into the relevant category. To improve the classification, we combine multiple noisy and imperfect models using weak supervision. This method, pioneered by Stanford AI, combines various labeling functions in an unsupervised probabilistic model, which learns the points of agreement and disagreement between functions to determine the most accurate label (Ratner et al., 2017).

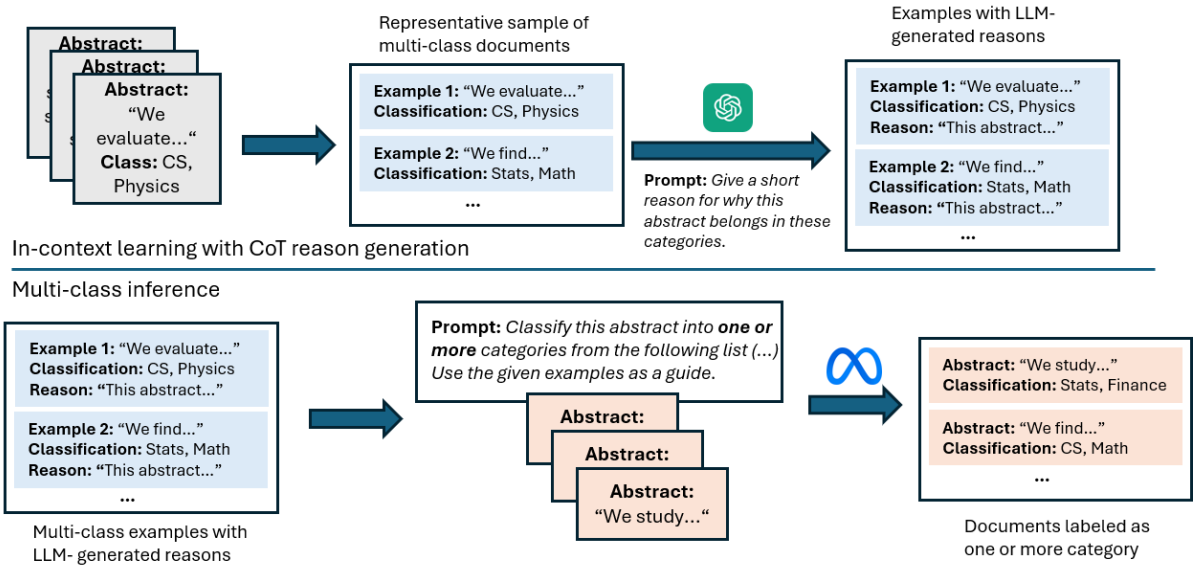

Figure 1: Overview of our proposed MultiICL approach. There are two stages: first, the generation of in-context learning examples with accompanying reasons, and second, inference with multi-class prompting. In the first stage, documents are randomly sampled from a set of labeled documents, and checked to ensure that they include sufficient overlap of all classes. An LLM is prompted to provide reasons for the classifications, which are added to the examples. Finally, the LLM is tasked with labeling a given abstract into one or more categories, using the examples from the first stage as a guide. The LLM then generates labels based on this information.

3.2 Overview of LLM based approach

In the LLM approach, we simply prompt the LLM to classify the document into a specific category. See the appendix for an example prompt for the field of computer science. We also experiment with in-context learning and chain-of-thought reasoning for prompts. Finally, we contrast two approaches to prompting: "binary prompting", where the LLM is tasked with answering if a given document does or does not belong in a category, and "multi-class prompting", where the LLM is given a list of labels and tasked with assigning a document to one or more of them.

3.3 In-context learning

In our approach, we leverage in-context learning (ICL) (Dong et al., 2022) as an effective solution to the data labeling problem. In-context learning is an emergent property of large language models (LLMs) where the model can perform a specific task using only a few examples within the prompt (Xie and Min, 2022), without requiring fine-tuning of its parameters. This approach can achieve high accuracy with minimal labeled data. In our experiments, we sampled examples from the data not included in the ablation or test sets, using two positive and three negative examples. This method already showed strong performance, but previous research indicates that optimizing the selection and format of prompts could further enhance results.

3.4 Chain-of-thought prompting

In our approach, we incorporate chain-of-thought prompting to improve the reasoning capabilities of large language models (LLMs) in data labeling tasks. Chain-of-thought refers to a series of intermediate reasoning steps that help LLMs perform better on reasoning-heavy tasks. We apply this technique by including explanations in our in-context examples, generated by GPT-3.5, and extend this by testing with newer models like Mistral 7B and GPT-4, along with a multi-class prompting method. Additionally, we conduct an ablation study using human-created explanations to guide the model away from common misclassifications, such as confusing mathematics articles with computer science articles. These explanations help refine the model's decision-making process and improve labeling accuracy.

3.5 Multi-class prompting and MultiICL

As far as we know, we are the first study to experiment with data labeling for a multi-class dataset. This allows us to experiment with two separate methods for classification and data labeling: splitting this task into separate binary classification

problems, and treating the dataset as a unified whole and performing multi-label classification, which we use on the phase 2 experiments.

Our method for performing multi-label classification uses a method we call "multi-class prompting", where the LLM is given the list of classes and asked to identify to which classes the article belongs. When combined with in-context learning, we call this method "MultiICL". Similar to previous experiments, we sample from data outside of those used in the experiments. Since we need to include examples that correspond to all six classes, we pick some examples that include combinations of different classes. To reduce the length of the prompt, we also remove definitions of each class from the prompt. The aim is to give the LLM a comprehensive understanding of the dataset, allowing it to develop a clearer intuition about how different classes intersect and where they remain distinct.

3.6 Weak supervision and LLMs

One experiment in the ablation study uses the powerful weak supervision framework Snorkel with LLMs treated as noisy sources of supervision. In this experiment, we treat the outputs of three smaller, weak LLMs as individual labeling functions, and use the Snorkel framework to produce better labels by having it observe and analyze where these LLMs agree and disagree.

4 Results

In this section, we describe our experimental set-up and summarize our results from 15 experiments.

4.1 Dataset

Our dataset consists of 20,972 research articles, each with titles, abstracts, and human-labeled classifications across six fields: computer science, physics, math, statistics, quantitative biology, and quantitative finance. This is a multi-class dataset, meaning each article can belong to one or more fields. However, the distribution is uneven: 41% of articles are in computer science, while only 1% are in quantitative finance. To address this, we create a balanced test set for phase 2 experiments, ensuring equal representation of each class. The dataset, sourced from Kaggle using arXiv papers, is initially tested through an ablation study on 200 articles using various LLMs and text embedding models to classify computer science articles. The best-performing models are then evaluated on a larger test set of 1,054 articles, with at least 200 items from each class to assess classification performance across all fields.

4.2 Experimental set-up

We run two sets of experiments that use a variety of different classification techniques using large language models and a weak supervision semantic similarity approach. This means that for each experiment, our goal is to validate the classification method by classifying some subset of documents and seeing how well it performs for different classes. We approach this multi-label classification task by either combining multiple binary classification models or using the multi-class prompting approach outlined in the methodology.

We use the following LLMs: Llama 3 8B, Llama 3 70B, GPT-3.5, GPT-4, and Mistral 7B (OpenAI et al., 2024) (AI@Meta, 2024) (Jiang et al., 2023). Llama 3 8B and Mistral 7B were run locally on Ollama, while GPT-3.5 and GPT-4 were run on using OpenAI's API. [1] Llama 3 70B was run on AI cloud inference platform Groq using an API. Each experiment took roughly 4-5 GPU hours, and consists of a single run.

We use the Massive Text Embedding Benchmark (MTEB) to choose embedding models (Muennighoff et al., 2023). The ablation study consists of moderately strong embedding models that are small enough to be run locally, while the phase 2 experiments use some of the strongest, largest models. A full list of embedding models used can be found in the appendix. Finally, we use the Snorkel framework to combine these embedding models (as well as the three small LLMs) in a weak supervision model (Ratner et al., 2017).

4.3 Evaluation metrics

We use four standard metrics to evaluate classification performance: accuracy, precision, recall, and F1-score. Accuracy reflects how many points were labeled correctly (the number of true positives and negatives as a percentage of the whole dataset). Precision refers to positive predictive value: of the positively labeled documents, how many were true positives? Recall is the opposite: of the true positives, how many were labeled as positive? The F1-score is the harmonic mean of precision and

[1] Licenses: Llama 3 (Community License), Mistral (MNPL), MTEB (Apache License). All model use complies with intended use.

recall; it is especially useful when dealing with an uneven class distribution, such as the phase 2 test set (Hossin and Sulaiman, 2015).

In both phases of the experiments, we place special emphasis on the accuracy and F1-score. The F1-score is especially important in the second phase of the experiments, where we deal with an uneven dataset where each class only accounts for 20-25% of the full test set. Accuracy and F1-score metrics are presented within the paper; see the appendix for the other performance metrics.

Category	Model	Accuracy	F1-Score
WS	Short def.	0.66	0.40
	Long def.	0.76	0.68
LLM	Llama 3 8B	0.68	0.70
	Llama 3 70B	0.90	0.87
	Mistral 7B	0.47	0.58
	GPT-3.5	0.63	0.67
	GPT-4	0.86	0.83
	LLMs + WS	0.78	0.76
LLM+ICL	Llama 3 8B	0.84	0.82
	Llama 3 70B	0.89	0.84

Table 1: Performance metrics of ablation study models

Class	WS	LLM	LLM+MC	LLM+ICL	MultiICL
CS	0.70	0.78	0.82	0.78	0.82
Phys	0.85	0.86	0.92	0.95	0.94
Math	0.76	0.72	0.84	0.90	0.77
Stats	0.78	0.72	0.78	0.69	0.73
QB	0.72	0.68	0.82	0.94	0.92
QF	0.84	0.92	0.96	0.95	0.97

Table 2: Accuracy metrics of Phase 2 models

Class	WS	LLM	LLM+MC	LLM+ICL	MultiICL
CS	0.50	0.70	0.71	0.70	0.71
Phys	0.56	0.70	0.80	0.86	0.82
Math	0.21	0.55	0.64	0.73	0.56
Stats	0.48	0.57	0.57	0.53	0.62
QB	0.51	0.52	0.65	0.84	0.78
QF	0.68	0.81	0.88	0.84	0.90

Table 3: F1-score metrics of Phase 2 models

4.4 Ablation study

Table 1 shows our ablation study results for the 200-item set, split into three sections: weak supervision (with the semantic similarity approach), LLMs with simple prompts, and LLMs with in-context learning. The model "LLMs + WS" refers to the combination of the three smallest models (GPT-3.5, Mistral 7B, Llama 3 8B) in a weak supervision model.

This first phase of experimentation reveals a number of notable insights. The best performing model is Llama 3 70B with simple prompting. Interestingly, this performs better than the same model with in-context learning. Similar results identifying equal or lesser performance with LLMs for in-context learning has been observed for GPT-3.5, with the reason perhaps being that instruction-tuning for these LLMs have decreased their in-context learning ability (He et al., 2023) (Fu and Khot, 2022). This model achieves a 1-point increase in accuracy and a 3-point increase in recall compared to the in-context learning version. Notably, it also outperforms GPT-4, with a 4-point increase in accuracy and a 3-point increase in recall.

Among smaller models, Llama 3 8B is the best-performing model, outperforming Mistral 7B by 21 points and GPT-3.5 by 5 points. Notably, when all three models are treated as labeling functions in a weak supervision model, there are significant gains in performance, with a 10-point increase in accuracy and 6-point increase in F1-score (compared to Llama 8B). This result shows how weak supervision can boost the performance of multiple small LLMs.

Unlike with the larger models, in-context learning can achieve notable gains in performance, putting smaller models on almost equal footing with larger ones. With in-context learning, Llama 3 8B achieves a 16-point increase in accuracy, along with 12-point increase in F1-score. In-context learning tends to decrease recall, which is what we see here with a 5-point decrease in recall; however, it makes up for it with larger gains in precision and accuracy.

On the weak supervision with semantic similarity side, we see that regardless of whether short or longer definitions are used, this method is less accurate than most LLMs. It is competitive with Llama 3 8B and even outperforms GPT-3.5 and Mistral 7B, although the lower recall may make it a worse choice.

4.5 Phase 2 Results

From this ablation study, we choose Llama 3 70B with and without in-context learning, along with the weak supervision approach with stronger embedding models, to move forward with the test set.

We also experiment with the multi-class prompting method for Llama 3 70B, again with/without in-context learning. Tables 2 and 3 show the accuracy and F1-score metrics from the five experiments on six classes, respectively. Precision and recall metrics can be found in the appendix (Table 9 and 10).

Comparing regular Llama 3 70B with a simple multi-class prompt change, we see improvements due to the multi-class prompt in every class, with an average accuracy increase of 7.7 points and F1-score increase of 6.7 points, with the largest increase for quantitative biology and mathematics.

Unlike the ablation study, major improvements are also seen when in-context learning is used with Llama 3 70B, although this varies between classes. We an average accuracy increase of 7.2 points, and an average F1-score increase of 10.8 points. Like with the multi-class method, the largest increases are for quantitative biology and mathematics. Given these results, we might expect that Llama 3 70B with multi-class prompting and in-context learning will perform the best. The results are somewhat mixed, but this variation, which we call "MultiICL", is slightly outperformed by the binary prompting method on average. MultiICL achieves an average accuracy increase of 6.17 points and an F1-score increase of 9 points. MultiICL outperforms regular ICL by 4 points in computer science and statistics for accuracy, but is worse in mathematics by 13 points (Table 2 and 3).

MultiICL achieves similar accuracy and F1-scores as regular ICL while requiring only a sixth of total prompts to the LLM, which means that it is both relatively inexpensive and takes less time. It achieves these high scores despite having half as many positive labeled examples in its prompt, and having no definitions. Given that the performance of in-context learning is also highly dependent on the prompts used, it is possible that MultiICL could perform even better than it has already with better examples or more examples. Given these considerations, MultiICL seems like the best solution for data labeling with multi-class data. When including labeled examples is difficult, regular multi-class prompting is also shown to be more effective than binary prompting with large models.

5 Discussion

The experiments presented above reveal a number of notable conclusions, one of which seems to be that the weak supervision with semantic similarity approach performs badly and is outperformed by LLMs across the board. This is not true: while it is indeed outperformed by LLMs, it is somewhat competitive against smaller models for physics (85% accuracy) and quantitative finance (84% accuracy), and even outperforms the LLMs for statistics (78%). What can explain the disparity in accuracy for these classes and the others?

One hypothesis is that the definitions are simply better than those of the other classes. Similarly, bad prompts could also explain the accuracy difference between computer science and statistics (which performed poorly) and the other fields for in-context learning. However, an analysis of how different field labels overlap in the dataset ("overlap analysis") is explained below to provide an alternative explanation for the discrepancies we see here.

It is plausible that weak supervision will perform better on some tasks more than others. However, achieving strong performance requires crafting the right definition and ensuring that the task is well-suited for weak supervision, something which is not entirely clear. In contrast, LLMs offer fairly strong performance, especially with in-context learning and prompt engineering.

	CS	Phys	Math	Stats	QB	QF
CS	276	16	37	98	16	9
Phys	16	201	10	6	0	0
Math	37	10	204	38	0	1
Stats	98	6	38	249	34	21
QB	16	0	0	34	200	4
QF	9	0	1	21	4	200

Table 4: Overlap table for true labels

	CS	Phys	Math	Stats	QB	QF
CS	455	64	163	236	214	70
Phys	64	308	142	70	173	24
Math	163	142	465	233	238	141
Stats	236	70	233	427	241	161
QB	214	173	238	241	505	84
QF	70	24	141	161	84	223

Table 5: Overlap table for Llama 3 70B

5.1 Overlap analysis

We now extend our analysis of automated data labeling methods by focusing on the multi-class prompting technique. Table 4 shows the true distribution of classes in the test set and how they overlap. Table 6 shows the overlap of predicted classes for Llama 3 70B with multi-class prompting

	CS	Phys	Math	Stats	QB	QF
CS	378	31	69	111	64	13
Phys	31	236	48	9	67	3
Math	69	48	269	53	49	6
Stats	111	9	53	293	51	48
QB	64	67	49	51	333	43
QF	13	3	6	48	43	166

Table 6: Overlap table for Llama 3 70B MC

	CS	Phys	Math	Stats	QB	QF
CS	365	11	38	235	40	18
Phys	11	168	80	5	14	2
Math	38	80	351	53	65	11
Stats	235	5	53	492	30	136
QB	40	14	65	30	198	2
QF	18	2	11	136	2	177

Table 7: Overlap table for Llama 3 70B MC + ICL

("Multi"). Table 7 shows the overlap of predicted classes for MultiICL. We contrast these numbers with Llama 3 70B with ICL prompting, which can be seen in Table 5. All tables can be found in the appendix.

When individual data points are better understood as the combination of classes, rather than individual ones, multi-class prompting is a useful technique. When comparing Llama 3 70B and Llama 3 70B MC, we see that while both tend to inflate the number of articles from each category, the numbers from Table 5 are generally much higher than those of Table 6 (see the diagonal on each table). The simple change in prompt results in numbers closer to the true labels. We hypothesize that the model seems to achieve greater intuition about the label composition of articles from knowing all the labels.

With regular multi-class prompting, the model is given no information as to the distribution of classes in the dataset. However, when in-context learning is added, the model may extrapolate the distribution of the dataset from the examples provided (see appendix for full MultiICL prompt). For example, MultiICL greatly inflates the number of articles that are both CS and Statistics, as well as those that are from Quantitative Finance and Statistics. Two of the examples in the ICL prompt happen to feature these characteristics, which suggests that the model has used the overlap of classes in the examples to infer the overlap of classes among the whole test set. (See the appendix for the full prompt)

The reverse phenomenon also seems to exist, as the model learns from the lack of overlap between classes as well. In the prompt, an abstract that only belongs to quantitative biology is included. Consequently, the number of quantitative biology articles decreases, as well as the overlap between it and other classes. Similar phenomena can be observed for the physics class, although interestingly the mathematics class does not adhere to this (Table 6 and 7).

These results underscore how multi-class prompting is a very powerful technique, but should be used with caution alongside in-context learning. The distribution of labeled examples can influence how the model labels its articles, so it is crucial that the distribution is representative. Including more examples in the prompt also seems like a promising solution to this issue.

6 Conclusion

In this paper, we evaluate two automated data labeling solutions: weak supervision using semantic similarity and large language models with and without in-context learning (ICL). Our findings show that the combination of multi-class prompting and ICL, which we call MultiICL, is the most effective, achieving high accuracy and F1 scores across six fields while requiring six times fewer prompts than methods without multi-class prompting. Though we experimented with a weakly-supervised NLP approach, it consistently under-performed compared to LLMs, likely due to the dependence on precise definitions. However, MultiICL significantly boosts accuracy by 7.70% over standard LLMs and is more efficient. This highlights the effectiveness of multi-class prompting in multi-label datasets, offering an efficient and powerful solution for data labeling. In the future, we will integrate additional context-aware features to enhance the accuracy and adaptability of MultiICL across a broader range of datasets. Additionally, we will systematically investigate the impacts of various types of bias in LLMs on model performance, ensuring the development of fairer and more equitable data labeling solutions.

7 Limitations

One limitation of our study is the use of a single dataset of research articles. While this dataset allowed us to perform a detailed comparison of data labeling methods, it may not capture the nuances of other contexts where different text structures or domain-specific language are present. In the future,

we will address this by testing our approaches on a wider variety of datasets, including those from diverse domains such as social media, legal documents, and technical reports. This will provide us with a broader perspective on the effectiveness of our methods across different data types.

Moreover, the performance of ICL methods is sensitive to the selection, ordering, and wording of examples, which can introduce variability and affect the reproducibility of results. This variability highlights the need for robust uncertainty quantification in ICL methods. In the future, we will explore techniques for quantifying uncertainty in model outputs, especially when different examples are used or when slight changes in prompts lead to varying predictions. We will also investigate automated example selection methods and adaptive frameworks that optimize both the presentation of examples and the estimation of uncertainty in real time. This approach will help improve the consistency, reliability, and transparency of ICL performance. Ultimately, by understanding and quantifying the uncertainty associated with example choices, we can develop more robust models suited for tasks requiring high reliability and precision.

Additionally, the approach of using LLMs for tasks like multi-class data labeling can introduce bias and fairness issues, as these models tend to replicate and even amplify existing biases found in their training datasets. This underscores the importance of incorporating human-annotated examples into in-context learning. By integrating carefully curated and diverse human annotations, we can guide the model more effectively, helping to mitigate biases and enhance the fairness and accuracy of the LLM's outputs. This human oversight is crucial for refining the model's performance and ensuring that the data labeling process remains as objective and equitable as possible.

References

AI@Meta. 2024. Llama 3 model card.

Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. 2020. Interactive weak supervision: Learning useful heuristics for data labeling. *arXiv preprint arXiv:2012.06046*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Michael Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Aaron N Richter, and Hamzah Al Najada. 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2:1–24.

Marika Cusick, Prakash Adekkanattu, Thomas R Campion Jr, Evan T Sholle, Annie Myers, Samprit Banerjee, George Alexopoulos, Yanshan Wang, and Jyotishman Pathak. 2021. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *Journal of psychiatric research*, 136:95–102.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Emad Elwany, Allison Hegel, Marina Shah, Brendan Roof, Genevieve Peaslee, and Quentin Rivet. 2022. Deeperdive: The unreasonable effectiveness of weak supervision in document understanding a case study in collaboration with uipath inc. *arXiv preprint arXiv:2208.08000*.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. 2017. Machine learning for medical imaging. *radiographics*, 37(2):505–515.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2020. Trove: Ontology-driven weak supervision for medical entity classification. *arXiv preprint arXiv:2008.01972*.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,

Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Alexander Kinsora, Kate Barron, Qiaozhu Mei, and VG Vinod Vydiswaran. 2017. Creating a labeled dataset for medical misinformation in health forums. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 456–461. IEEE.

Tanwi Mallick, Joshua David Bergerson, Duane R Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. 2023. Analyzing the impact of climate change on critical infrastructure from the scientific literature: A weakly supervised nlp approach. *arXiv preprint arXiv:2302.01887*.

Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham. 2008. A practical approach to classify evolving data streams: Training with limited amount of labeled data. In *2008 Eighth IEEE International Conference on Data Mining*, pages 929–934.

Niluthpol Chowdhury Mithun, Rameswar Panda, and Amit K Roy-Chowdhury. 2016. Generating diverse image datasets with limited labeling. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 566–570.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on*, pages 38–48. PMLR.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

OpenAI, Josh Achiam, Steven Adler, and et.al. Sandhini Agarwal. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.

Ryan Smith. 2023. Using few-shot learning language models as weak supervision.

Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM/JMS Journal of Data Science*, 1(2):1–30.

Refuel Team. Llms can label data as well as humans, but 100x faster.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sang Michael Xie and Sewon Min. 2022. How does in-context learning work? a framework for understanding the differences from traditional supervised learning.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Appendix

A.1 Additional performance metrics

Our analysis focused largely on the accuracy and F1-score metrics, which are presented in the paper. In the following tables, we show precision and recall metrics for the ablation study and phase 2 experiments.

Category	Model	Precision	Recall
WS	Short def.	0.64	0.29
	Long def.	0.70	0.65
LLM	Llama 3 8B	0.55	0.96
	Llama 3 70B	0.82	0.94
	Mistral 7B	0.42	0.95
	GPT-3.5	0.51	0.95
	GPT-4	0.76	0.91
	LLMs + WS	0.65	0.92
LLM+ICL	Llama 3 8B	0.74	0.91
	Llama 3 70B	0.82	0.91

Table 8: Precision and recall of ablation study models

Class	WS	LLM	LLM+MC	LLM+ICL	MultiICL
CS	0.45	0.55	0.61	0.55	0.62
Phys	0.63	0.58	0.74	0.87	0.90
Math	0.30	0.40	0.57	0.79	0.44
Stats	0.55	0.45	0.53	0.41	0.46
QB	0.38	0.37	0.52	0.80	0.79
QF	0.55	0.77	0.97	0.94	0.96

Table 9: Precision metrics of Phase 2 models

Class	WS	LLM	LLM+MC	LLM+ICL	MultiICL
CS	0.57	0.91	0.84	0.92	0.82
Phys	0.51	0.89	0.87	0.85	0.75
Math	0.16	0.91	0.75	0.67	0.76
Stats	0.43	0.78	0.62	0.72	0.92
QB	0.78	0.93	0.86	0.88	0.78
QF	0.89	0.86	0.81	0.76	0.85

Table 10: Recall metrics of Phase 2 models

A.2 Selected GPT-generated long definitions

In two of the phase 2 experiments, we use GPT-4 to generate definitions of fields, then compare it to the abstract to compute the semantic similarity and classify the abstract. Below is a selection of GPT-generated long definitions (as opposed to short definitions, which were also generated).

> Computer Science is a multifaceted discipline that investigates the design, analysis, and application of algorithms and computational systems to process and manage information efficiently. It encompasses core theoretical areas such as algorithm design, computational complexity, and cryptography, alongside applied domains like artificial intelligence, machine learning, software engineering, and cybersecurity. Research in computer science also extends to data science, human-computer interaction, and the development of innovative technologies for networks, databases, and embedded systems. This field is fundamental to advancing technology and solving complex problems across various sectors, from healthcare and finance to robotics and social media.

> Physics is the scientific study of the fundamental principles governing matter, energy, and their interactions. It explores theoretical domains such as quantum mechanics, relativity, and particle physics, while also delving into applied fields like condensed matter physics, electromagnetism, optics, and astrophysics. Research in physics seeks to understand everything from the smallest subatomic particles to the largest cosmic structures, uncovering the laws of nature and driving technological advancements. Through both experimental and theoretical approaches, physics aims to explain and predict the behavior of the universe at all scales, contributing to innovations across numerous fields, from electronics and materials science to healthcare and space exploration.

A.3 Embedding models

We use embedding models to generate vectors that represent the semantic content of a given document, then compute the cosine similarity between vectors. Below is a list of the specific models used in both experiment phases:

For ablation study experiments, with moderately strong models: UAE-Large-V1, GIST-large-Embedding-v0, bge-large-zh-v1.5, all-MiniLM-L6-v2, gte-base-en-v1.5, mxbai-embed-large-v1.

For phase 2 experiments, with the strongest models on the Massive Text Embedding Benchmark: voyage-lite-02-instruct (Rank 2), SFR-Embedding-Mistral (Rank 4), mxbai-embed-large-v1 (Rank 6), MUG-B-1.6 (Rank 7), Linq-Embed-Mistral (Rank 8), multilingual-e5-large-instruct (Rank 12)

A.4 Example prompt: MultiICL

Below is the prompt used for MultiICL in the phase 2 experiments. Note that this prompt is shortened. The actual prompt would include the full abstracts and reasons of five examples. Please refer to the attached Jupyter notebook corresponding to this experiment for the full prompt.

Task:
Classify this abstract into one or more categories from the following list: Computer Science (1), Physics (2), Mathematics (3), Statistics (4), Quantitative Biology (5), Quantitative Finance (6). Each field has a number associated with it, shown in parentheses. Please respond with a number or list of numbers, like "1,2" or "4" that correspond to the fields to which you believe the research article belongs. An article can belong to multiple fields, but please only assign the article to a field if it is explicitly focused on topics from that field. Please only respond with the number(s), with nothing else. Use the given examples as a guide:

Example 1:

- Abstract: "Using a large-scale Deep Learning approach..."
- Classification: 4,6
- Reason: "This abstract belongs to the fields of statistics because it employs non-parametric evidence and predictive modeling techniques, and quantitative finance due to its focus on analyzing market price dynamics, order flow, and stock prediction using large-scale financial data."

Example 2:

- Abstract: "Polycrystalline diamond coatings..."
- Classification: 2
- Reason: "This abstract belongs..."

Please classify this abstract:

A.5 Example prompt: Multi-class prompt

Below is the prompt used for the regular multi-class prompt (without in-context learning examples) in the phase 2 experiments.

Task:
I will paste an abstract from a research article, and I want you to classify it into one or more categories from the following list: Computer Science (1), Physics (2), Mathematics (3), Statistics (4), Quantitative Biology (5), Quantitative Finance (6). Each field has a number associated with it, shown in parentheses. Please respond with a number or list of numbers, like "1,2" or "4" that correspond to the fields to which you believe the research article belongs. An article can belong to multiple fields, but please only assign the article to a field if it is explicitly focused on topics from that field. Please only respond with the number(s), with nothing else. Please classify this abstract:

A.6 Example prompt: Binary prompt

Below is an example of the "binary prompt," in which an LLM is tasked with simply deciding whether an abstract belongs in a given category or not; this process is then applied for all of the categories to construct a multi-label dataset.

Task:
I will paste an abstract from a research article. I want you to respond "1" if the abstract comes from a paper in the computer science field, and "0" if it doesn't or if you are not sure. Do not include anything else in your response; just the number. Please classify this abstract: