ZeroDL: Zero-shot Distribution Learning for Text Clustering via Large Language Models

Hwiyeol Jo[1], **Hyunwoo Lee**[2], **Kang Min Yoo**[1], **Taiwoo Park**[3*]
[1]NAVER Cloud, [2]NAVER, [3]NAVER Search US
{hwiyeol.jo,hanu.lee,kangmin.yoo,taiwoo.park}@navercorp.com

Abstract

The advancements in large language models (LLMs) have brought significant progress in NLP tasks. However, if a task cannot be fully described in prompts, the models could fail to carry out the task. In this paper, we propose a simple yet effective method to contextualize a task toward a LLM. The method utilizes (1) open-ended zero-shot inference from the entire dataset, (2) aggregate the inference results, and (3) finally incorporate the aggregated meta-information for the actual task. We show the effectiveness in text clustering tasks, empowering LLMs to perform text-to-text-based clustering and leading to improvements on several datasets. Furthermore, we explore the generated class labels for clustering, showing how the LLM understands the task through data.

1 Introduction

Large language models (LLMs) have demonstrated impressive performances on various downstream tasks (Devlin et al., 2019; Radford et al., 2019). These also exhibit the ability to understand the context of input text, known as in-context learning (ICL) (Brown et al., 2020; OpenAI, 2023). ICL allows leveraging LLMs for specific tasks without further extensive training. However, effective use of ICL hinges on well-designed prompts.

While prompts with few-shot examples demonstrably improve performance, they can easily overfit a model to the examples (Perez et al. (2021); Mizrahi et al. (2023); *inter alia*). This led to a growing interest in zero-shot learning, which reduces the need for intricate few-shot selection. Recent advancements in zero-shot learning involve incorporating more sophisticated use of prompt structures, such as Chain-of-Thought (Wei et al., 2022), zero-shot reasoning (Kojima et al., 2022), and models trained to follow instructions (Ouyang et al., 2022; Chung et al., 2024). However, how to design prompts for target tasks remains challenging.

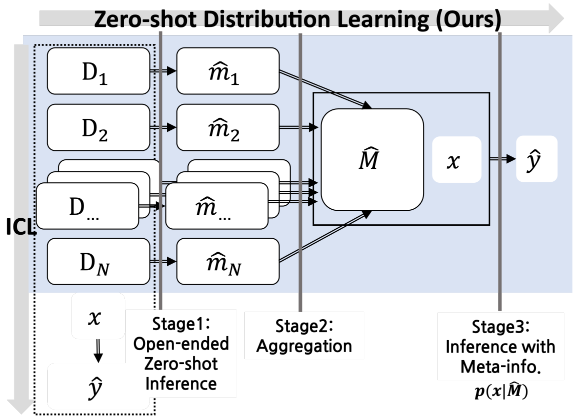

Figure 1: Illustration of proposed method ZeroDL. While in-context learning (ICL) relies on examples (D) tailored to specific tasks, ZeroDL aggregates all the outputs from these zero-shot inferences ($\hat{m}$), resulting in meta-level information ($\hat{M}$). This information is then used by the LLM to generate its final predictions.

Motivated by the core principle of ICL–providing task and data contexts *within* prompts–we propose an approach to construct more effective zero-shot prompts by understanding how LLMs describe datasets *across* prompting outputs.

As illustrated in Figure 1, **Zero**-shot **D**istribution **L**earning (ZeroDL) aims to learn data distributions through zero-shot inferences. The method comprises two key components: open-ended zero-shot inference and output aggregation. Zero-shot prompts are then constructed with the generated meta information, and used for actual task. This method takes advantage of the self-generated frame of LLMs to successfully carry out a given task.

We exemplify the effectiveness of ZeroDL on text clustering tasks where complete task descriptions cannot be provided due to an absence of ground-truth class labels. In addition, our method works in a text-to-text format, allowing clustering with specific context. For instance, "I love this movie" and "I hate this movie" express opposite sentiment but belong to the same cluster of movie reviews (see Table 9 in the Appendix for potential

[*] Now at Google

Findings of the Association for Computational Linguistics: ACL 2025, pages 19597–19607
July 27 - August 1, 2025

risk of absence of the perspectives).

Our contributions in this paper are as follows:

- We propose a novel approach called **Zero**-shot **D**istribution **L**earning that leverages zero-shot inferences to generate meta-level information about the data distribution by aggregating open-ended inference outputs from datasets.
- ZeroDL allows models to perform text-based clustering, empowering them to handle data with specific context, which offers advantages over embedding-based clustering methods.
- ZeroDL is competitive against embedding-based clustering methods on several datasets. Notably, ZeroDL even achieves better performances than models with ground-truth class labels in some cases.

2 Related Works

The majority of existing LLM-based clustering approaches rely on traditional methods like K-Means, employing LLM-generated embeddings as input (Petukhova et al., 2024; BehnamGhader et al., 2024). ClusterLLM (Zhang et al., 2023) introduced LLM-guided refinement of embedding-based clustering models. In contrast, our approach ZeroDL, directly leverages text-level prompting to inject specific viewpoints into LLMs, enabling more targeted and contextualized clustering.

Other works have explored using LLMs as clustering models (Wang et al., 2023; Viswanathan et al., 2023; Pham et al., 2023; Huang and He, 2024), but these approaches typically relied on few-shot settings with access to gold labels. IDAS (De Raedt et al., 2023) addressed this limitation by selecting representative data using an auxiliary embedding model to generate pseudo-labels without demonstrations. These labels are subsequently refined through in-context learning.

ZeroDL distinguishes itself by operating in a fully zero-shot setting, eliminating the need for additional embedding models. Moreover, while IDAS's class labels are primarily influenced by a limited set of data points and generate the labels independently, ZeroDL considers the entire data distribution through an aggregation step. This enables our method to generate class labels in an auto-regressive manner, taking into account the relationships between different labels.

The importance of appropriate ground-truth labels extends beyond clustering and permeates ICL. While Min et al. (2022) observed cases where the input-label correspondence does not play significant roles, Yoo et al. (2022) argued that the impact heavily depends on target tasks and experiment settings. We believe that this work would serve as a reference that appropriate class labels can be successfully generated by LLMs themselves.

3 Proposed Method: ZeroDL

Stage 1: Open-Ended Inference We begin by designing a prompt for zero-shot classification. This prompt intentionally avoids any detailed information about the task, minimizing the risk of overfitting. Based on the idea, we opt for the simplest prompt format:

Open-Ended Inference Prompt Template

Text: `[text]`

Classify the text to the best `[type_of_task]` class.

where `[text]` is the input data and `[type_of_task]` provides view of the task. In the experiment, it can be either sentiment or topic. Leveraging this prompt, we perform model inferences on all the input data (D). This process generates open-ended class predictions, which will be denoted as $\hat{m}$.

Stage 2: Aggregation The open-ended predictions lack constraints, leading to potentially inconsistent output formats. For instance, the model might predict "positive" and "non-positive" classes, while the ground-truth is "positive" and "negative". The predictions could even be entire sentences. To address the inconsistencies in the open-ended predictions, we employ aggregation strategy.

Before the aggregation, we count the frequency of each predictions and sort it, denoted as $\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_U$ where U denotes the unique number of predictions and the frequency of $\hat{m}_n$ is equal or larger than $\hat{m}_{n+1}$. After that, the predictions which frequency is only 1 are dropped in order to remove extraordinary predictions and save computation. Next, we iteratively construct subsets of the predictions by removing the least frequent predictions one by one. This process results in a list of subsets, denoted as S; $S_U = \{\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_U\}$, $S_{U-1} = \{\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_{U-1}\}$, $\cdots$, $S_1 = \{\hat{m}_1\}$. We input each subset to the LLM, providing more weights to the frequently occurred predictions (e.g., $\hat{m}_1$ is repeatedly included in the subsets while $\hat{m}_U$ occurs only once). The prompt template for aggregation is as follows:

Aggregation Prompt Template

[type_of_task] List:

S_U	S_{U-1}	$S_{\dots}$	S_1
$\hat{m}_1$	$\hat{m}_1$	$\hat{m}_1$	$\hat{m}_1$
$\hat{m}_2$	$\hat{m}_2$	$\hat{m}_2$	
$\dots$	$\dots$	$\dots$	
$\hat{m}_{U-1}$	$\hat{m}_{U-1}$		
$\hat{m}_U$			

Aggregate the [type_of_task] List into [NUM_CLUSTER_CLASS] classes.

where [NUM_CLUSTER_CLASS] is pre-defined number of classes to cluster. However, the model outputs often still lack coherence, especially in generating exact number of classes. To address this, we select the generation results only when the number of aggregated classes matches the pre-defined number of cluster classes. We then use the most frequent aggregated classes as meta-information $\hat{M}$ (class labels in the experiments). This information represents the model's understanding of potential views over the entire data.

Stage 3: Leveraging Meta-Information We incorporate the aggregated meta-information ($\hat{M}$) into the original prompt[1]:

Final Prediction Prompt Template

Text: [text]

Class description:
- Class 0: $\hat{M}_0$
- Class $\cdots$: $\hat{M}_{\dots}$
- Class C: $\hat{M}_C$

Based on the class description, classify the text to the best [type_of_task] class.

By incorporating the meta-information ($\hat{M}$) into the prompt, we enable the LLM to perform conditioned classification within the clustering context.

4 Experiments

Models Our primary experiments utilize mistral-7b-instruct-v0.2 (Jiang et al., 2023). We further evaluate our approach with several other models, including Qwen-2.5 (Team, 2024), Gemma-2 (Team et al., 2024), Llama-3.1 (Meta, 2024). Lastly, HyperCLOVA X SEED (Yoo et al., 2024) demonstrates the result of a relatively small, instruct-following language model. All the results are averaged over 5 runs.

[1]The order of Text: and Class description: can be reversed.

Baselines IDAS (De Raedt et al., 2023) represents a zero-shot clustering approach. We employ 8 demonstrations as described in the paper, using SBERT (Opitz and Frank, 2022) as an additional embedder. Final prediction is performed using the same prompt template as our Stage 3. We also include llm2vec (BehnamGhader et al., 2024) with K-Means where applicable. This method represents a baseline using (not exactly but) the same backbone computed by embedding-level[2]. We exclude other few-shot settings from our comparisons as they are not considered fair due to the inherent advantages of leveraging additional information from demonstrations and the subjectivity involved in the few-shot selection.

Setting Our method aims to effectively cluster data distributions, converging them into a suitable number of classes. We recognize that clustering datasets often exhibit a high number of classes relative to the available data points. To address this, we utilize text classification datasets, which typically offer a larger number of data instances per class than clustering datasets. Moreover, the availability of ground-truth labels facilitates direct comparison between generated and actual labels, enabling qualitative analysis and interpretability.

Datasets We thus use 6 text classification datasets for clustering: IMDB (Maas et al., 2011), SST-2, SST-5 (Socher et al., 2013), YelpReivews (Zhang et al., 2015) for sentiment classification, and AGNews, DBpedia (Lehmann et al., 2015), YahooAnswers (Chang et al., 2008) for topic classification. Data information is presented in Table 4 of the Appendix.

Evaluation Model performance is evaluated based on the highest accuracy achieved across all possible mapping combinations. To determine the predicted class, we leverage the `Class n` anchor tokens within the LLM outputs. However, LLMs might not directly predict the same classes as the ground-truth labels[3], so we test all possible mapping combinations and report the best performed one. We split datasets with more than 7 classes (i.e., DBpedia and YahooAnswers) into 2 subsets Front (F) and Back (B) to avoid out-of-memory[4] after

[2]Note that IDAS requires an additional embedding model and llm2vec involves additional training steps.

[3]For example, `Class 0` means `Positive` in the prediction while `Negative` is labeled as `Class 0` in the ground-truth.

[4]The computation is a factorial of the number of classes.

Model	Method	IMDB	SST-2	SST-5	YRev	AGNew	DBp(F)	DBp(B)	Yah(F)	Yah(B)	Macro	Micro
hcx-seed -0.5b-it	ZeroDL(C-T)	56.2	62.5	33.6	**43.6**	**33.1**	24.1	16.9	**21.3**	**18.3**	34.4	33.1
	IDAS(C-T)	88.3	74.9	27.7	35.0	29.4	17.0	26.8	-	-	-	-
	Gold(C-T)	**72.1**	**75.2**	**35.2**	27.6	31.4	31.2	24.1	18.3	**18.3**	**37.0**	**33.6**
	ZeroDL(T-C)	86.7	72.2	**36.0**	**43.9**	**59.5**	32.3	44.5	26.7	22.8	47.2	46.4
	IDAS(T-C)	**89.4**	62.9	28.9	28.7	27.3	23.7	27.2	-	-	-	-
	Gold(T-C)	89.1	**73.7**	34.4	39.1	59.4	**61.1**	**57.2**	**40.3**	**56.2**	**56.7**	**56.3**
mistral -7b-it	llm2vec+KMeans	62.1	55.2	30.3	**56.0**	**84.4**	**96.1**	70.8	**48.0**	46.2	61.0	67.5
	ZeroDL(C-T)	**88.8**	**85.7**	40.5	48.7	75.2	58.9	64.1	47.2	61.4	63.4	61.5
	IDAS(C-T)	51.1	63.1	41.4	37.3	38.1	32.1	21.8	30.2	41.8	39.7	35.4
	Gold(C-T)	87.5	77.0	**41.8**	50.8	60.7	74.2	**85.2**	40.8	**62.5**	**64.5**	**68.1**
	ZeroDL(T-C)	90.2	84.2	36.0	46.8	79.5	56.7	72.2	**51.0**	66.3	64.8	63.0
	IDAS(T-C)	50.1	68.1	39.6	41.0	44.6	35.8	27.6	33.3	46.1	42.9	38.9
	Gold(T-C)	**91.7**	**82.5**	**43.3**	51.8	82.7	84.1	**82.7**	50.9	**73.8**	**71.5**	**72.7**
Qwen-2.5 -7b-it	ZeroDL(C-T)	87.6	74.8	**45.5**	**53.8**	**70.9**	62.3	75.9	45.3	**68.8**	65.0	65.8
	IDAS(C-T)	**93.3**	**87.3**	42.7	51.1	68.8	47.4	68.4	30.3	41.8	58.9	59.0
	Gold(C-T)	84.3	**87.4**	45.0	49.9	62.2	**84.5**	**89.0**	**51.9**	68.2	**69.2**	**71.3**
	ZeroDL(T-C)	93.5	85.2	**50.5**	49.3	**84.3**	66.7	78.3	50.8	74.4	70.3	68.2
	IDAS(T-C)	**95.3**	**88.9**	45.1	**51.6**	83.1	56.0	72.2	31.6	43.3	63.0	62.7
	Gold(T-C)	94.7	88.5	46.3	49.0	82.9	**94.2**	**98.0**	**64.5**	**79.7**	**77.5**	**78.6**
gemma-2 -27b-it	ZeroDL(C-T)	94.3	88.0	**52.6**	51.6	**74.5**	52.8	53.5	50.3	60.0	64.2	60.1
	IDAS(C-T)	**95.3**	**91.1**	42.7	**58.3**	31.7	56.9	44.0	-	-	-	-
	Gold(C-T)	79.8	86.3	48.7	57.3	70.5	**89.8**	**97.1**	**56.1**	**65.9**	**72.4**	**75.9**
	ZeroDL(T-C)	94.0	88.3	50.8	50.8	81.8	57.3	57.6	52.7	73.0	67.4	62.7
	IDAS(T-C)	**95.2**	89.2	42.4	**55.6**	35.0	60.5	49.8	-	-	-	-
	Gold(T-C)	95.3	**89.6**	**51.8**	**55.7**	**85.4**	**92.6**	**98.7**	**66.9**	**84.4**	**80.0**	**81.0**

Table 1: The performance of ZeroDL for text clustering. C-T denotes the prompt order with class information then input text. T-C is the reversed. Bold means the best accuracy in the same prompt order and underline denotes the outperforming cases than baselines (except for Gold label). IDAS with HyperCLOVA X SEED and Gemma-2 is constrained by the maximum 8,192 token length. More baselines are presented in Table 5 of the Appendix.

removing the 3 smallest size of classes[5].

5 Results

Table 1 presents the performance of our ZeroDL method. Notably, ZeroDL surpasses models provided with ground-truth (Gold) class labels on several datasets. This suggests that ZeroDL may uncover richer or more nuanced class labels within the data compared with pre-defined labels. Our method demonstrates performance comparable to K-Means clustering using LLM embeddings, particularly excelling on datasets with relatively smaller sizes. ZeroDL outperforms IDAS, especially on datasets with more than 2 classes. IDAS often struggles to generate appropriate class labels, particularly when relying on a limited coverage of representative data instances as demonstrations. Furthermore, IDAS is susceptible to the input length limitations of LLMs when the demonstrations are long, as evident in the Gemma results. In contrast, ZeroDL achieves its results through flexible zero-shot prompting without requiring any modifications.

Table 2 shows examples of class labels generated from ZeroDL. These labels often provide richer and more informative explanations[6] compared with the original ground-truth labels. Moreover, ZeroDL can potentially uncover novel classes based on the data, such as Ambiguous Sentiment and Mixed Sentiment[7]. This highlights the limitations of pre-defined labels in capturing the full spectrum of sentiment complexity.

Figure 2 illustrates the confusion matrix of the results. The relatively low performance of SST-5, even with gold labels, underscores the inherent difficulty of sentiment classification. In contrast, AG-News demonstrates successful matching with our method. This observation emphasizes the potential value of ZeroDL in scenarios where ground-truth labels are unavailable.

Table 3 investigates the significance of class la-

[5] CarsAndTransportation, SocialScience, and Sports classes in YahooAnswers dataset.

[6] Note that the generation of class description varies depending on datasets and LLMs.

[7] Examples are provided in Table 6 of the Appendix.

Data	Method	ClassLabels
SST-5	ZeroDL	**Neutral Sentiment**: This class includes all the sentiment labels that express a neutral sentiment towards the movie or documentary. Examples include (...) **Negative Sentiment**: This class includes all the sentiment labels that express a negative or sad emotion towards the movie or documentary. Examples include (...) **Ambiguous Sentiment**: This class includes all the sentiment labels that do not clearly express a positive or negative emotion towards the movie or documentary. Examples include (...) **Mixed Sentiment**: This class includes all the sentiment labels that express a mixed sentiment towards the movie or documentary. Examples include (...) **Positive Sentiment**: This class includes all the sentiment labels that express a positive emotion towards the movie or documentary. Examples include (...)
	IDAS	**Disappointing**, **Critical**, **Negative**, **Positive**, **Disappointing or Unsatisfying**
	Gold	**Very Negative**, **Negative**, **Neutral**, **Positive**, **Very Positive**
AGNews	ZeroDL	**International Relations and Politics**: This class includes topics related to international relations, diplomacy, Middle East politics, terrorism, nuclear politics, and elections. **Sports and Entertainment**: This class includes topics related to sports, tennis, golf, basketball, baseball, cricket, and entertainment. **Business and Economy**: This class includes topics related to the economy, finance, stocks, mergers and acquisitions, retail, real estate, and labor markets. **Technology and Science**: This class includes topics related to technology, computing, internet, cybersecurity, space exploration, and science.
	IDAS	**Iran's Nuclear Program, Oil Prices, Businesses prioritize hardware upgrades in economic recovery, Volcanic activity at Mount St. Helens**
	Gold	**World**, **Sports**, **Business**, **Sci/Tech**

Table 2: The example of generated class labels in 5-class sentiment classification (`SST-5`), and topic classification (`AGNews`). ZeroDL can generates alternative class labels and its description. Compared with IDAS that has overlap in the generated labels, ZeroDL considers other labels in generation. Furthermore, IDAS's representative data points fail to generate general topic class labels. Additional examples are in Table 8.

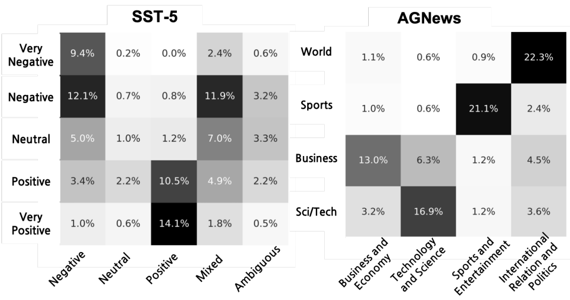

Figure 2: Confusion matrix of proposed method ZeroDL in SST-5 (left) and AGNews (right). x-axis and y-axis are generated labels and gold labels, respectively.

	S2(C-T)	S2(T-C)	S5(C-T)	S5(T-C)
RandToken	51.6	59.5	28.6	28.6
AutoL(Best)	**86.8**	**84.9**	**43.9**	40.5
AutoL(Worst)	82.5	80.7	41.1	39.7
Gold	77.0	82.5	41.8	**43.3**
ZeroDL(Mistral)	80.6	83.7	40.3	40.5
ZeroDL(gpt-4.1)	77.0	82.5	41.0	40.3

Table 3: The mistral-7b performance with various class labels. `ZeroDL(Mistral)` uses only the generated class titles for fair comparison. The best labels from `AutoL` in SST-2 are [Wonderful, Bad] and the worst are [Irresistible, Pathetic]. In SST-5, [Terrible, Better, Good, Extraordinary, Unforgettable] are the best while [Awful, Better, Hilarious, Perfect, Wonderful] are the worst. `gpt-4.1-mini` generates [Positive, Negative] (the same as Gold) and [Positive, Negative, Neutral, Mixed, Other / Unclear], respectively.

bels in text clustering tasks. We explore the performance of `mistral-7b-instruct-v0.2` using class labels suggested by AutoL (Gao et al., 2021) designed for prompt-based model fine-tuning. These labels represent a curated selection. We also investigate the class labels generated by `gpt-4.1-mini`. The results demonstrate that clustering performance is generally higher when using manually selected labels than using the original dataset labels. This suggests that carefully chosen labels can significantly improve clustering outcomes. In the context, ZeroDL performs particularly well on SST-5, implying that the potential of ZeroDL to capitalize on informative class labels automatically.

ZeroDL involves a trade-off in computational cost. We evaluate the impact of varying input data amounts on performance (see Table 7 in the Appendix). While using only 10% of the data yields reasonable performance, it can lead to inconsistencies in model performance as evidenced by increased standard deviation.

6 Conclusion

We introduce ZeroDL, a novel approach to contextualize tasks for a given LLM. ZeroDL employs open-ended zero-shot inference and output aggregation to learn data distributions. We demonstrate its effectiveness, showing competitive performances against embedding-based clustering methods and superior performance than ground-truth labels in some cases. Beyond its clustering capabilities, ZeroDL offers the generation of informative class labels that provide deeper insights into LLMs.

7 Limitations

Prompt Dependency and Heuristics ZeroDL relies on carefully designed prompts to guide LLMs towards effective clustering. While our focus was on using simple and intuitive prompts, prompt selection can potentially influence the model's behavior and introduce biases. Future work could explore more sophisticated prompt engineering techniques to further enhance ZeroDL's performance.

Experiments with Diverse LLMs and Prompts While we acknowledge the computational limitations (and price) of ZeroDL, investigating its behavior with a wider range of LLMs (including commercial models like GPT, Claude, and Gemini) and prompt templates could provide valuable insights into the generalizability and robustness of the approach.

Lower than State-of-the-Art Performance Achieving state-of-the-art performance is not the sole focus of ZeroDL but it offers a valuable framework in understanding data distributions with zero-shot inference via LLMs. By addressing the limitations mentioned above, ZeroDL has the potential to become a powerful and versatile tool not only for text clustering but data exploration.

Expensive Computational Cost in Inferences. Although ZeroDL have an alternative approach to reduce computational burden through data sampling, effectively sampling data to generate appropriate class labels remains a challenge. Although we report the trade-off in computational cost and performance in Table 7 of the Appendix, the technique within the framework presents a valuable future direction.

Acknowledgement

The authors would like to thank reviewers in several rounds of discussion. Thanks to Alice Lee for her help with English writing and Minji Hong for her support in writing this work. Lastly, we extend our gratitude to Hyperscale AI Team at NAVER Cloud, for unwavering support in connecting us with key resources throughout this research.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*, volume 2, pages 830–835.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. Idas: Intent discovery with abstractive summarization. *arXiv preprint arXiv:2305.19783*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanjun Gao, Ting-Hao Huang, and Rebecca J. Passonneau. 2021. Learning clause representation from dependency-anchor graph for connective prediction. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 54–66, Mexico City, Mexico. Association for Computational Linguistics.

Chen Huang and Guoxiu He. 2024. Text clustering as classification with llms. *arXiv preprint arXiv:2410.00927*.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *arXiv preprint arXiv:2401.00595*.

OpenAI. 2023. Gpt-4 technical report.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. 2024. Text clustering with llm embeddings. *arXiv preprint arXiv:2403.15112*.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. *arXiv preprint arXiv:2305.13749*.

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.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. *arXiv preprint arXiv:2305.14871*.

A Appendix

A.1 Data Information

	#Train	#Valid	#Test	#Class	Pre-defined ClassTitle
IMDB	25,000	3,750	25,000	2	Negative, Positive
SST-2	9,645	1,101	2,210	2	Negative, Positive
SST-5	9,645	1,101	2,210	5	Very Negative, Negative, Neutral, Positive, Very Positive
YelpReviews	650,000	97,500	49,999	5	Very Negative, Negative, Neutral, Positive, Very Positive
AGNews	120,000	18,000	7,600	4	World, Sports, Business, Sci/Tech
DBpedia(F)	280,000	41,939	35,000	7	Company, EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTransportation, Building
DBpedia(B)	280,000	42,213	35,000	7	NaturalPlace, Village, Animal, Plant, Album, Film, WrittenWork
Yahoo(F)	59,518	8,879	10,489	7	ArtsAndHumanities, BeautyAndStyle, BusinessAndFinance, ComputersAndInternet, ConsumerElectronics, EducationAndReference, EntertainmentAndMusic
Yahoo(B)	59,493	8,973	10,514	7	FoodAndDrink, GamesAndRecreation, Health, HomeAndGarden, Pets, PregnancyAndParenting, SocietyAndCulture

Table 4: Data information used in the experiments.

A.2 Additional Model Performances

Model	Method	IMDB	SST-2	SST-5	YRev	AGNew	DBp(F)	DBp(B)	Yah(F)	Yah(B)	Macro	Micro
TF-IDF	KMeans	52.1	53.0	26.2	35.2	43.8	55.9	62.8	44.7	46.5	46.7	48.8
SBERT	KMeans	65.0	54.0	33.9	30.3	82.7	95.4	93.9	67.2	68.3	65.6	67.5
llama-3.1-8b-it	ZeroDL(C-T)	76.0	**73.8**	35.0	**51.1**	54.5	37.7	58.5	46.4	47.9	53.4	53.2
	IDAS(C-T)	**92.2**	**73.8**	42.4	45.4	55.4	40.8	64.1	29.2	35.3	53.2	53.9
	Gold(C-T)	80.4	73.5	**44.6**	48.7	**64.1**	**74.4**	**80.8**	**49.6**	**69.5**	**65.1**	**66.7**
	ZeroDL(T-C)	94.4	79.4	39.1	49.4	77.6	46.1	68.1	53.7	60.0	63.1	61.1
	IDAS(T-C)	**94.6**	**84.5**	38.4	**50.1**	67.7	37.6	66.4	32.0	36.2	56.4	56.2
	Gold(T-C)	94.4	81.5	**43.4**	**50.1**	**78.9**	**80.8**	**75.1**	**62.8**	**78.3**	**71.7**	**71.2**
Qwen-2.5-72b-it	ZeroDL(C-T)	94.5	**90.2**	44.5	**62.6**	71.3	60.8	64.4	53.4	71.3	68.1	67.5
	IDAS(C-T)	**95.3**	88.8	47.1	50.5	68.1	62.3	57.2	35.7	55.8	62.3	61.1
	Gold(C-T)	95.1	89.4	**51.0**	57.7	**81.3**	**94.2**	**97.9**	**69.0**	**84.8**	**80.0**	**81.6**
	ZeroDL(T-C)	95.3	**90.1**	45.4	**63.3**	76.6	65.2	65.4	55.1	73.6	70.0	69.4
	IDAS(T-C)	**95.9**	89.9	46.3	51.2	76.4	63.2	58.9	34.5	47.7	62.7	61.7
	Gold(T-C)	**95.8**	**90.0**	**52.7**	59.9	**83.9**	**96.8**	**99.7**	**73.7**	**87.4**	**82.2**	**83.8**

Table 5: The performance of ZeroDL for text clustering compared with more methods. C-T denotes the prompt order with class information then input text. T-C is the reversed. Bold means the best accuracy in the same prompt order and underline the outperforming cases than baseline except for ground-truth class labels.

A.3 Qualitative Examples

Predicted	Gold	Example
Mixed Sentiment	VeryNeg	It 's hard not to feel you 've just watched a feature-length video game with some really heavy back story .
	Neg	But it pays a price for its intricate intellectual gamesmanship .
	Neutral	The appearance of Treebeard and Gollum 's expanded role will either have you loving what you 're seeing, or rolling your eyes .
	Pos	An utterly compelling ' who wrote it ' in which the reputation of the most famous author who ever lived comes into question .
	VeryPos	... a roller-coaster ride of a movie
Ambiguous Sentiment	VeryNeg	It 's difficult to say whether The Tuxedo is more boring or embarrassing – I 'm prepared to call it a draw .
	Neg	Like most Bond outings in recent years , some of the stunts are so outlandish that they border on being cartoonlike .
	Neutral	Effective but too-tepid biopic
	Pos	But he somehow pulls it off .
	VeryPos	Emerges as something rare , an issue movie that 's so honest and keenly observed that it does n't feel like one .

Table 6: Example of data predicted to newly generated classes in SST-5. Randomly selected.

A.4 Data Sampling Ablation

[C-T]	IMDB	SST-2	SST-5	YRev	AGNew	DBp(F)	DBp(L)	Yah(F)	Yah(L)
All	88.8	85.7	40.5	48.7	75.2	58.9	64.1	47.2	61.4
(std)	3.20	0.18	1.22	0.21	1.22	7.85	6.61	1.02	1.92
1%	81.3	79.3	46.9	51.5	66.0	59.4	58.4	36.1	43.7
(std)	2.97	1.5	0.88	0.12	5.71	1.93	5.11	0.72	3.35
5%	74.6	83.6	45.8	49.4	74.2	58.8	66.4	39.3	58.5
(std)	18.49	1.37	2.78	0.47	1.94	5.38	9.13	4.82	2.73
10%	87.0	84.5	44.8	49.2	74.5	67.2	63.9	40.4	60.4
(std)	5.90	1.50	5.02	0.29	1.44	6.35	8.95	5.11	2.96
[T-C]	IMDB	SST-2	SST-5	YRev	AGNew	DBp(F)	DBp(L)	Yah(F)	Yah(L)
All	90.2	84.2	36.0	46.8	79.5	56.7	72.2	51.0	66.3
(std)	3.52	0.24	3.84	0.47	1.73	9.73	6.86	0.75	2.80
1%	88.4	58.6	42.0	49.8	74.6	57.7	67.5	37.7	49.4
(std)	0.51	11.93	3.99	2.05	4.34	1.31	8.37	4.1	3.15
5%	85.5	80.2	41.7	46.2	80.7	57.1	74.1	37.2	68.1
(std)	7.36	1.92	4.09	1.42	2.83	10.58	9.36	6.87	4.58
10%	91.2	82.0	39.1	46.6	79.9	63.2	75.3	45.2	68.5
(std)	2.16	1.95	2.26	0.33	2.74	6.53	9.91	3.43	3.42

Table 7: ZeroDL performances according to the number of input data.

A.5 Additional Examples of Generated Classes

Data	Method	ClassLabels
IMDB	ZeroDL	**Negative Sentiment**: The list also includes various expressions of negative sentiment towards movies, films, shows, and documentaries. Some examples include (...) **Positive Sentiment**: The list includes various expressions of positive sentiment towards movies, films, shows, and documentaries. Some examples include (...)
	Gold	**Negative**, **Positive**
SST-2	ZeroDL	**Positive Sentiment**: All the sentiment labels that express a positive sentiment towards the movie, film, documentary, or subject. For example, (...) **Negative or Neutral Sentiment**: All the sentiment labels that do not express a positive sentiment towards the movie, film, documentary, or subject. For example, (...)
	Gold	**Negative**, **Positive**
Yelp Rev	ZeroDL	**Negative Sentiment**: This class includes sentences expressing negative sentiments towards a place, food, or experience. Examples include (...) **Very Positive Sentiment**: This class includes sentences expressing highly positive sentiments towards a place, food, or experience. Examples include (...) **Mixed Sentiment**: This class includes sentences expressing mixed sentiments towards a place, food, or experience. Examples include (...) **Neutral Sentiment**: This class includes sentences expressing neutral sentiments towards a place, food, or experience. Examples include (...) **Positive Sentiment**: This class includes sentences expressing positive sentiments towards a place, food, or experience. Examples include (...)
	Gold	**Very Negative**, **Negative**, **Neutral**, **Positive**, **Very Positive**
DBp (F)	ZeroDL	**Aviation and Transportation**: This class includes topics related to aviation, aerospace technology, military history, maritime history, and transportation. **Business and Economy**: This class includes topics related to business, finance, industries, companies, and economics. **Sports and Biographies**: This class includes topics related to sports, athletes, and their biographies. **Politics and Government**: This class includes topics related to politics, government, elections, and specific political parties. **Education**: This class includes topics related to education, universities, schools, and specific educational institutions. **History and Architecture**: This class includes topics related to history, architecture, historic sites, castles, and landmarks. **Art and Entertainment**: This class includes topics related to art, music, entertainment, and specific artists or record labels.
	Gold	**Company**, **EducationalInstitution**, **Artist**, **Athlete**, **OfficeHolder**, **MeanOfTransportation**, **Building**
DBp (B)	ZeroDL	**Science and Technology**: This class includes topics related to paleontology, geology, volcanology, space exploration, and academic journals. **Music and Entertainment**: This class includes topics related to music, album releases, jazz music, heavy metal music, hip hop music, and entertainment. **Geography and Hydrology**: This class includes topics related to geography, hydrology, rivers, water bodies, and water resources. **Botany and Plant Sciences**: This class includes topics related to botany, horticulture, plant taxonomy, plant conservation, and endangered species. **Literature and Books**: This class includes topics related to literature, novels, fiction, mystery and crime fiction, and academic publications. **Zoology and Entomology**: This class includes topics related to zoology, entomology, moths, butterflies, fish species, and arachnids. **Film and Television**: This class includes topics related to film, cinema, movies, movie reviews, Bollywood, and television.
	Gold	**NaturalPlace**, **Village**, **Animal**, **Plant**, **Album**, **Film**, **WrittenWork**
Yah (F)	ZeroDL	**Personal Finance and Economics**: Topics related to personal finance, credit scores, debt management, taxes, and economics. **Health and Wellness**: Topics related to health, medicine, fitness, nutrition, and wellness. **Pop Culture and Entertainment**: Topics related to music, movies, TV shows, books, art, and entertainment. **Technology and Computing**: Topics related to computers, technology, software, internet, telecommunications, and mobile phones. **Miscellaneous**: Topics that do not fit neatly into any of the above categories, such as philosophy, religion, science, and humor. **Fashion and Beauty**: Topics related to fashion, clothing, makeup, cosmetics, hair care, and beauty. **Education and Careers**: Topics related to education, academic programs, scholarships, student loans, careers, and employment.
	Gold	**ArtsAndHumanities**, **BeautyAndStyle**, **BusinessAndFinance**, **ComputersAndInternet**, **ConsumerElectronics**, **EducationAndReference**, **EntertainmnentAndMusic**
Yah (B)	ZeroDL	**Food and Cooking**: This class includes topics related to various cuisines, recipes, food items, and cooking techniques. **Home Improvement and DIY**: This class includes topics related to home repair, renovation, decorating, gardening, and DIY projects. **Miscellaneous**: This class includes topics that do not fit neatly into any of the above categories, such as politics, education, art, and entertainment. **Technology and Gaming**: This class includes topics related to video games, computer hardware, software, technology, and internet culture. **Health and Wellness**: This class includes topics related to physical and mental health, nutrition, dieting, weight loss, fitness, exercise, and medical conditions. **Animals and Pets**: This class includes topics related to various animals, pet care, animal rights, and wildlife. **Religion and Philosophy**: This class includes topics related to various religions, theology, philosophy, and spirituality.
	Gold	**FoodAndDrink**, **GamesAndRecreation**, **Health**, **HomeAndGarden**, **Pets**, **PregnancyAndParenting**, **SocietyAndCulture**

Table 8: Additional examples of generated class labels. Class title is marked as bold and its description is colored as gray.

Data	TaskType	Generated Class Labels (Descriptions are omitted)
IMDB	sentiment	Negative, Very Positive, Highly Negative, Mixed, Neutral-Positive, Positive, Highly Positive, Neutral-Negative, Extremely Negative, Neutral; [Total 10]
	topic	Film and Television Influence on Business, Film and Television Influence on Politics, Film and Television Production Companies, Film and Television Influence on Art, Film and Television Productions, Film Analysis or Review, Film and Television Influence on Society and Culture, Film and Television Influence on Technology, Film and Television Recommendations, Film and Television Awards, Film and Television, Film and Television Genres, Film and Television Influence on Society, Film and Television Marketing, Movie Reviews or Film Criticism, Film and Television Critics, Film and Television Technologies, Film and Television Festivals, Film or Media Criticism, Film and Television History, Film and Television Technology Trends, Film and Television Education, Film and Television Influence on Entertainment, Film and Television Influence on Education, Personal Opinion or Review, Film and Television Industry, Film and Television Distribution; [Total 27]
SST	sentiment	The text expresses a negative sentiment towards the subject being described, The text expresses a negative or sad sentiment towards the film, The text expresses a negative or sad sentiment, The text has a negative tone, The text expresses a negative or cautionary sentiment, The text expresses a negative or slightly negative sentiment, Negative Sentiment, The text expresses a negative or slightly negative sentiment towards the film, The text expresses a negative sentiment, The text expresses a negative sentiment towards the movie being described, The text has a negative sentiment; [Total 11]
	topic	Literature and Writing, Media and Entertainment, Mental Health and Psychology, Travel and Adventure, Film and Television, Family and Relationships, Science and Technology, Food and Cooking, Performing Arts, Sports, Education and Learning, Business and Finance, Greetings and Open-Ended Texts, Entertainment; [Total 14]
Yelp Rev	sentiment	Positive Sentiment, Highly Negative Sentiment, Neutral to Positive Sentiment, Negative Sentiment, Mixed Sentiment, Neutral Sentiment, Very Positive Sentiment, Neutral to Negative Sentiment; [Total: 8]
	topic	Automotive, Environment or Nature, Legal or Law, Customer Service or Business, Nightlife or Entertainment, Education or Training, Sports or Fitness, Discrimination or Racism, Religion or Spirituality, Technology or Gadgets, Health or Medical, Travel or Tourism, Education or Learning, Beauty or Personal Care, Customer Reviews or Testimonials, Politics or Government, Food or Cooking, Food or Beverage, Shopping or Retail, Home Improvement or Construction, Science or Technology, Real Estate or Housing, Food Safety or Food Poisoning, Entertainment or Leisure, Arts or Culture, Personal Care or Beauty, Business or Economy, Personal Experiences, Dining Experience or Food Review; [Total 29]
AG News	sentiment	Negative: 12 expressions that contain negative sentiment, Positive: 25 expressions that contain positive sentiment, Negative: 12, Sentiment not clear: 3, Mixed: 11 expressions that contain a mix of positive and negative sentiment, Positive: 25, Neutral: 33, Neutral: 33 expressions that do not contain any clear positive or negative sentiment, Sentiment not clear: 3 expressions that do not provide enough context to determine a clear sentiment, Mixed: 11; [Total 10]
	topic	Same Result with Table 2
DBp (F)	sentiment	Neutral, Positive, Negative; [Total 3]
	topic	Baseball, Aviation, Sports and Biographies, History, Aircraft, Soccer or Football, People and Biographies, Ice Hockey, Education, American Football, General, Aircraft Design, Music or Entertainment, Higher Education or Universities, Healthcare or Hospitals, Football or Soccer; [Total 15]
DBp (B)	sentiment	Ambiguous, Neutral, Positive, Romantic, Negative, Mixed, Objective; [Total 6]
	topic	Botany or Biology (specifically, Plant Science or Taxonomy), Botany or Endangered Species or Conservation Biology, Botany or Horticulture, Botany or Algae, Botany or Brazilian Flora, Botany or Cacti, Botany or Mexican Flora, Botany or Tillandsia species, Botany or Orchids, Botany or Aquatic Plants, Botany or Plant Science, Botany or Tropical Plants, Botany or Hawaiian Flora, Botany or Plant Taxonomy, Botany or Palm Trees [Total 15]
Yah (F)	sentiment	Positive, Neutral, Mixed, Informational, Negative [Total 5]
	topic	Miscellaneous (for topics that do not fit neatly into any specific category), Education and Careers, Makeup and Beauty, Gaming and Technology, Fashion and Clothing, Telecommunications, Housing and Real Estate, Music and Entertainment, Pop Culture and Entertainment, Personal Finance and Credit Scores; [Total 10]
Yah (B)	sentiment	Neutral; [Total 1]
	topic	Health and Medical Concerns, Pop Culture and Entertainment, Mental Health and Psychology, Literature and Writing, Business and Finance, Food and Cooking, Video Games and Technology, Art and Creativity, Philosophy and Ethics, Education and Learning, Humor and Satire, Sports and Fitness, Religion and Theology, Home Improvement and DIY, Science and Technology, Travel and Adventure, Pets and Animals, Politics and Society, Pregnancy and Reproductive Health; [Total 19]

Table 9: The examples of generated class labels when no constraints are given; we experiment with different task type and unlimited the number of clusters.