Label Set Optimization via Activation Distribution Kurtosis for Zero-Shot Classification with Generative Models

Yue Li, Zhixue Zhao and Carolina Scarton
Department of Computer Science, University of Sheffield, UK
{yli381,zhixue.zhao,c.scarton}@sheffield.ac.uk

Abstract

In-context learning (ICL) performance is highly sensitive to prompt design, yet the impact of class label options (e.g. lexicon or order) in zero-shot classification remains underexplored. This study proposes LOADS (Label set Optimization via Activation Distribution kurtosiS), a post-hoc method for selecting optimal label sets in zero-shot ICL with large language models (LLMs). LOADS is built upon the observations in our empirical analysis, the first to systematically examine how label option design (i.e., lexical choice, order, and elaboration) impacts classification performance. This analysis shows that the lexical choice of the labels in the prompt (such as *agree* vs. *support* in stance classification) plays an important role in both model performance and model's sensitivity to the label order. A further investigation demonstrates that optimal label words tend to activate fewer outlier neurons in LLMs' feed-forward networks. LOADS then leverages kurtosis to measure the neuron activation distribution for label selection, requiring only a single forward pass without gradient propagation or labelled data. The LOADS-selected label words consistently demonstrate effectiveness for zero-shot ICL across classification tasks, datasets, models and languages, achieving maximum performance gain from 0.54 to 0.76 compared to the conventional approach of using original dataset label words.

1 Introduction

Generative large language models (LLMs) are increasingly used for classification tasks via zero-shot in-context learning (ICL), where models are prompted to select an option from a pre-defined set of labels (Wang et al., 2022; Antypas et al., 2023; Gonen et al., 2023; Mu et al., 2024). While some classification tasks employ a relatively fixed set of lexicons to represent class labels, such as sentiment analysis (*positive* and *negative*) and textual entailment (*entailment* and *contradiction*), other tasks may present more ambiguous choices in lexical selection. Stance classification, for instance, uses diverse pairs of antonyms to represent positive and negative stances across different datasets, e.g. *agree-disagree* vs. *favor-against*. As a result, when crafting prompts for these classification tasks, practitioners face decisions regarding label options in the prompt, such as lexical selection and ordering.

Despite studies suggesting the sensitivity of ICL to prompt design (Lu et al., 2022; Yoo et al., 2022; Wei et al., 2024; Mao et al., 2024; Liu et al., 2024a; Zhang et al., 2022; Liu et al., 2022b; Peng et al., 2024; Gonen et al., 2023; Mu et al., 2024), this subtle yet critical consideration of label options in prompt for zero-shot ICL has received limited attention. To fill in this research gap, we explore the impact of three types of label variants (i.e., lexical choice, label order and elaborations) in zero-shot ICL with both encoder-decoder and decoder-only LLMs. We mainly ground our research on stance classification, a task where label adaptation is a known problem due to various label inventories in different studies (Hardalov et al., 2021). We demonstrate that the lexical choice of the label options significantly impacts model performance. The model's sensitivity to the label order also depends on the lexical choice, while elaborating on task-related information (e.g. *agree with the claim* elaborating *agree*) has minimum effect.

Inspired by recent studies on neuron analysis (Kuzmin et al., 2023; Voita et al., 2024; Stolfo et al., 2024; Kurz et al., 2024), we further investigate the neurons in the feed-forward network (FFN) in the decoder of the LLMs. We empirically show that prompts with optimal label sets activate fewer outlier neurons. Consequently, we propose a new method, **L**abel set **O**ptimization via **A**ctivation **D**istribution kurtosi**S** (LOADS), to select optimal label sets for a given classification dataset in zero-shot ICL. LOADS could stably and effectively work with only 100 unlabelled samples of the val-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31724–31741
November 4-9, 2025 ©2025 Association for Computational Linguistics

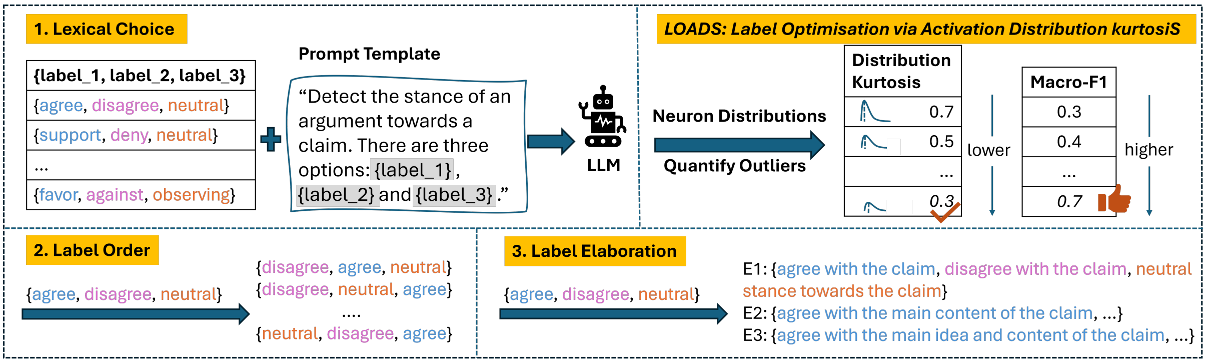

1. Lexical Choice
{label_1, label_2, label_3}
{agree, disagree, neutral}
{support, deny, neutral}
...
{favor, against, observing}
Prompt Template
"Detect the stance of an argument towards a claim. There are three options: {label_1}, {label_2} and {label_3}."
LLM
LOADS: Label Optimisation via Activation Distribution kurtosiS
Neuron Distributions
Quantify Outliers
Distribution Kurtosis
0.7
0.5
...
0.3
lower
Macro-F1
0.3
0.4
...
0.7
higher
2. Label Order
{agree, disagree, neutral}
{disagree, agree, neutral}
{disagree, neutral, agree}
....
{neutral, disagree, agree}
3. Label Elaboration
{agree, disagree, neutral}
E1: {agree with the claim, disagree with the claim, neutral stance towards the claim}
E2: {agree with the main content of the claim, ...}
E3: {agree with the main idea and content of the claim, ...}

Figure 1: Illustration of the three aspects (i.e., lexical choice, label order and label elaboration) for designing the label option in the prompt in zero-shot ICL for classification, and our LOADS to post-hoc select the optimal label set (top half figure).

idation dataset, also demonstrating transferability across datasets and languages. Our contributions are summarized as follows:

- The first benchmark on **how variants of label options in prompts affect zero-shot ICL models' performance** for classification tasks. We provide useful recommendations on label designing to practitioners working on zero-shot classification with LLMs [1].
- The empirical demonstration that **zero-shot ICL performance negatively correlates with the number of outlier neurons in FFN** when varying the lexical choices for label options. The correlation holds true across diverse English stance classification datasets and topic classification datasets with different models.
- A **novel and efficient post-hoc method (LOADS) for label selection in zero-shot ICL**. Compared with common strategies in practice, our approach demonstrates statistically significant performance improvements across model types, model sizes and languages with only 100 unlabeled data samples. Our analysis also suggests that the LOADS-selected label set is potentially transferable across similar datasets for a specific LLM, further alleviating the cost to collect samples for a target new dataset.

We present our experimental setups, results and discussions on the impact of label options in zero-shot ICL in Section 3. Then, we describe our neuron analysis of the lexical choice in label options in Section 4 and our proposed method LOADS for selecting optimal label sets in Section 5.

2 Related Work

ICL Performance Few-shot ICL mainly focuses on cases where LLMs are directly prompted with N demonstration examples. The studies highlight the substantial impact of example ordering (Lu et al., 2022), formatting (Yoo et al., 2022; Wei et al., 2024; Mao et al., 2024; Liu et al., 2024a), and examples selection (Zhang et al., 2022; Liu et al., 2022b; Peng et al., 2024). The parallel lines of work focus on improving few-shot ICL via the optimal selection or arrangement of examples (Liu et al., 2022b; Rubin et al., 2022; Lu et al., 2022; Zhang et al., 2022; Liu et al., 2024b; Xu et al., 2024), reweighting examples (Yang et al., 2023), automatic reformat or generation of demonstration representations (Kim et al., 2022; Liu et al., 2023), and introduction of intermediate reasoning steps (Wei et al., 2022; Zhang et al., 2023b). However, the impact of lexical choices for label names in classification received little attention, with the only closely related work suggesting that LLMs are likely to confuse classes which share similar key vectors in the attention modules (Wang et al., 2023).

For zero-shot ICL, Mu et al., (2024) demonstrate the effect of using synonyms for class options, but they neither consider the order of the label options nor propose an effective strategy to choose the label names. Gonen et al., (2023) empirically show that perplexity could be an effective indicator for prompt selection, but they do not account for class options. Notably, behavioral differences between few-shot and zero-shot ICL have been frequently observed, suggesting that findings from few-shot ICL do not necessarily hold in the zero-shot context (Lin and Lee, 2024).

[1] Code and resources: `https://github.com/YLi999/Stance_LOADS`.

Prompt-Tuning and Verbalizer Prompt-tuning aims to automatically find or generate an optimal discrete prompt (e.g., through gradient-based search (Shin et al., 2020; Shi et al., 2023) or fine-tuning (Gao et al., 2021; Le Scao and Rush, 2021; Deng et al., 2022)) or by training continuous prompt (Lester et al., 2021; Liu et al., 2022c). Verbalizer can be taken as a mapping function that links discrete class labels to corresponding tokens or phrases in a model's vocabulary. A range of methods developed to build the verbalizer, including manually created verbalizer (Schick and Schütze, 2021), search-based verbalizer that identifies label words automatically from the dataset (Gao et al., 2021; Shin et al., 2020), and soft verbalizers that uses continuous embeddings obtained through fine-tuning (Hambardzumyan et al., 2021; Cui et al., 2022). Prompt-tuning often does not focus on label set selection for zero-shot ICL, and the verbalizer introduces additional components to the decoding of the generative models, distinguishing it fundamentally from our work.

3 Prompting with Varied Label Options for Zero-shot Classification

In zero-shot ICL for classification, a common approach is to provide a set of class label options in the prompt to instruct the LLMs to choose one of the options as the classification prediction. Although the label option is a subtle component in the prompt, we are interested in whether it has a significant impact on model performance.

Specifically, we explore three types of variants around label options in the prompt: (1) *lexical choice*; (2) *label order*; and (3) *label elaboration*. To accurately measure the impact of these factors, we only manipulate the label options within the same prompt template (Section 3.2). We show examples of the three variants in Figure 1.

3.1 Methodology

Lexical Choice We use single-word synonyms to represent class labels (e.g. *support* and *agree*), forming various label sets. For each dataset, we compare the zero-shot ICL performance when LLMs are prompted to select from different label sets, as illustrated in Figure 1. For this purpose, we design a pipeline to create a pool of label sets:

1. *Collect a seed set of label names.* We obtain this set by collecting the label names in the datasets we experiment with.
2. *Expand label sets with WordNet and LLMs.* We use WordNet (Fellbaum, 1998) as a reliable source and Claude[2] as a supplementary source to obtain synonyms for label names in the seed set. For pairs of label names with semantically opposite meanings (such as "agree" and "disagree"), we also consider antonyms to avoid potential ambiguity and present clear contrast for the predicted models.
3. *Manual selection.* We manually filter out semantically unrelated or inappropriate label sets generated by Claude to mitigate the impact of noisy label names.

The label names are arranged in the sequence presented in their original study (see Table 1). We refer to this arrangement as the *default order*.

Label Order We consider every possible order of the single-word labels in the prompt and compare the model performance against that obtained with the default order. For binary datasets, there is only one alternative arrangement besides the default order, while N-way multi-class classification would yield $N! - 1$ alternative orders.

Label Elaboration We investigate whether transforming single-word labels (e.g., "agree") into more detailed phrases (e.g., "agree with the claim") has an impact on model performance. On the one hand, elaborating on task details with the label may provide the model with a stronger alignment signal between the label and the task, emphasizing what the label is referring to. On the other hand, it also increases the label length and may introduce noise in the prompt (Liu et al., 2024a). Therefore, we design three levels of elaborations (shorted for *E1*, *E2* and *E3*) by progressively adding more task-related (and potentially redundant) information to the single-word label names, as shown in Figure 1.

3.2 Experimental Setups

Datasets We focus on the stance classification task due to its rich label inventories across various readily available datasets. Stance classification aims to identify the type of an expressed opinion (e.g., "agree" or "disagree") in a given piece of text towards a particular topic, claim, or entity. We consider four binary (*scd* (Hasan and Ng, 2013), *perspectrum* (Chen et al., 2019), *snopes* (Hanselowski et al., 2019) and *ibmcs* (Bar-Haim et al., 2017)) and

[2] https://claude.ai/new

Dataset Name	Original Label Words	Optimal Label Words
scd	for, against	pro, con
perspectrum	support, undermine	validate, refute
snopes	agree, refute	affirm, refute
ibmcs	pro, con	endorse, deny
vast	pro, con, neutral	confirm, dispute, neither
emergent	for, against, observing	endorse, reject, neutral
semeval	favour, against, neither	accept, reject, neutral
rumoureval	support, deny, query, comment	confirm, reject, question, neutral
arc	agree, disagree, discuss, unrelated	affirm, refute, discuss, unrelated

Table 1: Lists of the English stance classification datasets, their labels in *original* dataset, and the *optimal labels* with the highest zero-shot ICL performance on Flan-T5-xl as an example to justify our motivation on LOADS.

five multi-class datasets (*vast* (Allaway and McKeown, 2020), *emergent* (Ferreira and Vlachos, 2016), *semeval* (Mohammad et al., 2016), *rumoureval* (Gorrell et al., 2019) and *arc* (Hanselowski et al., 2018)) from existing English stance classification benchmarks (Schiller et al., 2021; Hardalov et al., 2021; Chen et al., 2023), as shown in Table 1. The nine datasets cover different domains, such as social media posts, news articles and online debates forums. We also experiment with topic classification in Section 4 and 5 to demonstrate the generalizability of our findings to other NLP tasks.

Models We cover both encoder-decoder and decoder-only LLMs, and experiment with the prevalent open-sourced Flan-T5 (Chung et al., 2024), Llama 3 and Llama 3.1 (Dubey et al., 2024) model families as representatives for these two types of LLMs. We choose their moderate-sized instruction-tuned versions, Flan-T5-xl (3b), Llama -3-Instruct (8b) and Llama-3.1-Instruct (8b), as our primary models for investigation due to our hardware resources constraints and their decent zero-shot ICL performance (Aiyappa et al., 2024; Chung et al., 2024; Dubey et al., 2024). We conduct experiments with *Gemma-2-it (9b)* [3] and *Flan-T5-xxl* (13b) to further show the generalizability of LOADS in Section 5.

Prompt Template We refer to the prompt template used in the supervised fine-tuning of Flan-T5 and Llama (Wang et al., 2022; Chung et al., 2024; Dubey et al., 2024). We present the results with the following prompt template in this paper: *Given a [text1_name] and a [text2_name], detect the stance that the [text2_name] has towards the [text1_name]. There are {N} options: “{label$_0$}", “{label$_1$}, ... , and {label$_{N-1}$}". Now complete the following example. [text1_name]: {text1}. [text2_name]: {text2}.* We also test other templates and find the results are consistent (e.g., prompting with label explanations in Appendix E).

Evaluation We adopt macro-$F1$ for model performance evaluation to align with prior studies (Schiller et al., 2021; Hardalov et al., 2021; Chen et al., 2023). We use $wF2$[4] to account for data imbalance in rumoureval (Scarton et al., 2020).

Implementation Details To ensure reproducibility, we use greedy search for decoding[5]. In more than 95% cases, LLMs exactly follow the instruction and output the stance name within the required stance options. Therefore, we directly use the model generation as the predicted label without post-processing or mapping. We run experiments on the validation set of each dataset. See Appendix C for details on the label sets experimented with. We exclude the topic/entities-based (such as stance towards Obama) stance classification datasets (i.e. scd, semeval and vast) in label elaboration experiments to avoid unnecessary ambiguity or bias during elaboration (e.g., the text to be classified may refer to anything from policy to personal behavior about Obama, and elaborating *agree* to *agree with the opinion of Obama* could lead to biased prediction).

3.3 Results and Analysis

We first discuss the impact of lexical choice, label order, and label elaboration on zero-shot ICL for classification. Then we provide suggestions for practitioners in zero-shot ICL for classification.

Lexical Choice As shown in Table 2, performance varies across datasets and models solely due to changes in the label names within the prompt. The variations are more pronounced than those reported in previous studies (Mu et al., 2024). The gap between the highest and lowest performance exceeds 0.1 for all datasets and models, and surpasses 0.2 on more than half of the datasets. We observe that certain stance labels could potentially trigger biased predictions, leading to extremely low performance. For example, Flan-T5 tends to always output *support* when using *support*, *deny*, *neither* for the semeval dataset, and Llama 3 overly predicts *con* when prompted by the label set *pro*,

[3] https://huggingface.co/google/gemma-2-9b-it

[4] $wF2$ gives different weights for each stance: *deny* = *support* = 0.40, *query* = 0.15 and *comment* = 0.05.

[5] Potential impact of the decoding strategy can be found in the Appendix D.

	F_{score}	**perspectrum**	**ibmcs**	**snopes**	**scd**	**emergent**	**semeval**	**rumoureval**	**arc**	**vast**
Llama 3 (8b)	*max*	0.869	0.834	0.770	0.759	0.740	0.688	0.659	0.453	0.382
	min	0.738	0.592	0.540	0.639	0.387	0.544	0.286	0.249	0.173
	Original	0.799	0.679	0.656	0.709	0.467	0.643	0.521	0.364	0.219
	avg±var	0.824±0.001	0.756±0.003	0.666±0.003	0.713±0.001	0.547±0.006	0.624±0.001	0.514±0.008	0.329±0.002	0.279±0.002
Llama 3.1 (8b)	*max*	0.909	0.898	0.748	0.769	0.660	0.735	0.584	0.466	0.409
	min	0.376	0.300	0.435	0.629	0.166	0.456	0.320	0.252	0.175
	Original	0.809	0.759	0.660	0.749	0.494	0.676	0.480	0.426	0.237
	avg±var	0.832±0.009	0.805±0.011	0.668±0.003	0.737±0.001	0.496±0.013	0.652±0.003	0.488±0.003	0.387±0.002	0.282±0.002
Flan-T5-xl (3b)	*max*	0.940	0.960	0.809	0.776	0.743	0.706	0.761	0.685	0.496
	min	0.836	0.807	0.576	0.449	0.487	0.166	0.281	0.358	0.155
	Original	0.939	0.939	0.746	0.631	0.649	0.467	0.381	0.493	0.311
	avg±var	0.899±0.001	0.901±0.002	0.695±0.004	0.661±0.009	0.626±0.003	0.539±0.012	0.520±0.010	0.507±0.008	0.328±0.008

Table 2: The maximum (*max*), minimum (*min*), average (*avg*), variance (*var*) of the model performance across different label names in the prompt for each validation set. The performance of the original label set (*Original*) is also included, showing that they fail to reach the maximum performance LLMs could get. The extent of the gap between the maximum and minimum performances is represented using colors: *max*−*min*> 0.3, 0.2 <*max*−*min*< 0.3. The greater the variance, the greater the impact of label lexical choice, and the greater the potential utility of optimizing the label set.

con, *neutral* for the emergent dataset.

Label Order On average, we observe limited influence of label order (see the Appendix F for details). However, we are also interested in extremes of performance change caused by shifting label orders, particularly the maximum performance gain and drop, and whether the extreme gain or drop correlates with the lexicons used for the label names. Therefore, for each dataset, we first select the top-k[6] optimal and poor label sets based on their performance in the same order (i.e. the default order). We then examine the maximum increase or decrease of the performance after re-arranging the label orders for optimal and poor label sets, respectively.

We find that altering the order for the optimal label sets has the risk of high performance drops (e.g., even more than 0.2 on the binary classification snopes dataset, Figure 2a in Appendix). While performance improvements are possible, the gains are relatively limited (lower than 0.1 on all datasets). Conversely, re-arranging the order for the poorly performing label sets offers potential for substantial improvement (Figure 2b). This high improvements for certain label sets after re-ordering suggests that the poor performance may partly stem from the initial sub-optimal ordering.

Label Elaborations Similarly, we select the top-k[7] optimal and sub-optimal single-word label sets based on their performance, and examine the performance change after elaborations. We observe that the models are robust to the elaborations for either optimal or poor single-name label sets (full results in Table 9), indicating that adding task related details or increasing the label token lengths brings limited impact on average. However, we also observe relatively large performance change on certain datasets. Specifically, for the rumoureval dataset with Llama 3, we notice a performance drop larger than 0.2 when elaborating an optimal label set. Performance on the ibmcs dataset with Llama 3.1 could increase 0.2 when elaborating a poorly-performing label set.

3.4 Suggestions to Practitioners

Based on our results and analysis, we provide the following suggestions to practitioners in zero-shot ICL for classification:

1. Lexical designing for the label names should be considered as an important step in prompt engineering.
2. Single-word class label without elaboration on task information is able to achieve high performance in most cases, i.e. adding extra information does not yield better results.
3. If the practitioner has selected a set of optimal lexicons for label options based on a specific order, exploring alternative label orders can be redundant due to the limited performance gain brought from high computational costs. However, if a label set is chosen randomly, experimenting with different label orders may yield meaningful improvements (see Figure 2).

[6] Due to the exponential increasing of label order options and our limited computational resources, we set k=15 for binary classification; k=10 for three-way classification and k=2 for four-way classification

[7] k=15 for binary classification, k=30 for three/four-way classification.

Model	Stance Classificaion									Topic Classification	
	perspectrum	ibmcs	snopes	scd	emergent	semeval	rumoureval	arc	vast	TweetTopic	AG News
Llama 3 (8b)	-0.4921*	-0.3787*	-0.4359*	-0.4217*	-0.5781	-0.2618*	-0.3764*	-0.1662	-0.3639*	-0.4994*	-0.2447*
Llama 3.1 (8b)	-0.4103*	-0.3642*	-0.1874	-0.0686	-0.4310*	-0.2232*	-0.3944*	-0.1708*	-0.1208	-0.2196*	0.0200
Flan-T5-xl (3b)	-0.4476*	-0.3638*	-0.6014*	0.2353	-0.1089	-0.2881*	-0.1714*	-0.5742	0.1003	0.1587	0.0398

Table 3: Spearman correlation co-efficiency between model performance on validation set and kurtosis of neurons in the last layer. Mark with * when p value is lower than 0.05.

4 Neuron Analysis for Label Selection

Although our findings indicate the importance of label word selection for text classification in zero-shot ICL, current studies lack consideration of this factor. Therefore, we conduct empirical analysis to gain insights into the underlying mechanism of lexical choice for single-word label names.

We preliminarily explored related approaches discussed in Section 2, including prompt perplexity (Gonen et al., 2023) and model internal representation of label words (Wang et al., 2023), but they did not yield any significant correlation with zero-shot ICL model performance (see the Appendix J and K for details). Meanwhile, various studies have indicated the correlation between neuron activation pattern in FFN and model performance (Kuzmin et al., 2023; Tang et al., 2024; Stolfo et al., 2024; Wu et al., 2024). Inspired by the finding that the presence of outliers in neural networks is predictive of quantization and pruning performance for the layers of LLMs (Kuzmin et al., 2023), we establish a new hypothesis: the model performance influenced by label names is correlated with the number of outliers in the neurons within FFN in the decoder of the LLMs. We empirically validate our hypothesis on the nine stance classification datasets, as well as two topic classification datasets (AG News (Zhang et al., 2015) and TweetTopic (Antypas et al., 2022)) to show the generalizability to other NLP tasks.

Methodology For each FFN module in layer i in the decoder, it can be denoted as follows:

$$h^i = (\text{act\_fn}(\tilde{h}^i W_1^i) \otimes \tilde{h}^i W_3^i) \cdot W_2^i. \qquad (1)$$

where $\tilde{h}^i$ is the output hidden states from multi-head self-attention module. The activation function (act_fn) for Flan-T5 and Llama 3/3.1 is Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel, 2016) and Sigmoid Linear Unit (SiLU) (Hendrycks and Gimpel, 2016; Elfwing et al., 2018), respectively.

A *neuron* is defined as the linear transformation of each column in W_1^i followed by the activation function. Here, we study the last layer I's output of act_fn$(\tilde{h}^I W_1^I)$ (denoted as N_I) for the predicted first token of the label name in the model generation. Following Kuzmin et al., (2023), we measure the number of outliers over the neuron output distribution (N_I) through kurtosis, given by:

$$\text{Kurtosis}[N_I] = \frac{\mathbb{E}[(N_I - \mu)^4]}{(\mathbb{E}[(N_I - \mu)^2])^2} \qquad (2)$$

where μ is the mean of N_I. For each dataset, we average the kurtosis scores over the validation set for each candidate label set. We then calculate the Spearman correlation between model performance and the averaged kurtosis score.

Results Table 3 shows that, for most datasets, there is statistically significant negative correlation between model performances and kurtosis scores across models and activation functions, indicating that fewer outliers in the neurons of the final layers are associated with enhanced zero-shot ICL performance. This observation implies that the kurtosis score of neuron activation distribution in FFN of the last decoder layer of LLMs could potentially serve as an effective signal for selecting optimal label names in zero-shot classification.

5 LOADS: Label set Optimization via Activation Distribution kurtosiS

Motivated by the above observation that the fluctuated zero-shot ICL performance caused by different label names could be attributed to the number of outliers in neurons in the last layer of LLMs, we propose LOADS to obtain an optimal label set for a given classification task in a post-hoc setting.

5.1 Method

We design a three-step pipeline based on LOADS for automatic label selection in zero-shot ICL:

1. Create a list of candidate label sets for class options in the prompt (see Section 3). The label names in each set should follow the same order.
2. Rank the list of label options based on the kurtosis score of the neuron activation in FFN of

Model	Method	Stance Classification									Topic Classification	
		perspectrum	ibmcs	snopes	scd	emergent	semeval	rumoureval	arc	vast	TweetTopic	AG News
Llama 3 (8b)	LOADS	0.8431	0.7684	0.5516	0.6698	0.5870	0.6445	0.4097	0.3662	0.3190	0.7752	0.7660
	Original Label	0.8187	0.5485	0.4387	0.6504	0.3416	0.5565	0.3487	0.3130	0.2395	0.7742	0.7594
	Original + Verbalizer	0.8185	0.5485	0.4306	0.6578	0.3400	0.5576	0.3476	0.3131	0.2395	0.7742	0.7594
	Self-generated	0.6912	0.5802	0.3314	0.6607	0.3692	0.6032	0.2745	0.2837	0.3070	0.5851	0.6235
Llama 3.1 (8b)	LOADS	0.8789	0.8856	0.6212	0.7274	0.6403	0.6581	0.3642	0.3637	0.2458	0.7988	0.7306
	Original Label	0.8064	0.6783	0.4426	0.6983	0.4337	0.6492	0.3528	0.3784	0.2126	0.7909	0.7273
	Original + Verbalizer	0.8064	0.6783	0.4426	0.6983	0.4262	0.6448	0.3534	0.3838	0.2128	0.7909	0.7273
	Self-generated	0.6244	0.4918	0.5146	0.6798	0.2984	0.6421	0.3409	0.3373	0.2578	0.5956	0.6217
Gemma 2 (9b)	LOADS	0.9110	0.9247	0.6233	0.7595	0.5989	0.6784	0.4896	0.4858	0.3439	0.8032	0.8451
	Original Label	0.8966	0.8728	0.6454	0.7707	0.5704	0.6874	0.4778	0.4858	0.3383	0.8241	0.8316
	Original + Verbalizer	0.8966	0.8728	0.6424	0.7707	0.5523	0.6873	0.4689	0.4748	0.3383	0.8246	0.8322
	Self-generated	0.9017	0.9113	0.6656	0.7480	0.5833	0.6621	0.3549	0.4657	0.3178	0.6303	0.8215
Flan-T5-xl (3b)	LOADS	0.9334	0.9380	0.6881	0.5997	0.5813	0.4951	0.4759	0.4688	0.3857	0.8530	0.8556
	Original Label	0.9305	0.8971	0.7267	0.6341	0.5580	0.5697	0.3837	0.4628	0.3473	0.8071	0.9209
	Original + Verbalizer	0.9305	0.8953	0.7026	0.6507	0.5586	0.5712	0.3852	0.4613	0.3482	0.8091	0.9209
	Self-generated	0.7914	0.7384	0.3677	0.6974	0.3883	0.5954	0.2986	0.3343	0.3042	0.8501	0.7368
Flan-T5-xxl (13b)	LOADS	0.9428	0.9644	0.6905	0.6158	0.5938	0.5257	0.3852	0.6151	0.4218	0.7727	0.7742
	Original Label	0.9407	0.9630	0.7814	0.7598	0.5614	0.5697	0.2016	0.6065	0.3278	0.6643	0.9177
	Original + Verbalizer	0.9407	0.9621	0.7716	0.7598	0.5631	0.5705	0.2011	0.6062	0.3278	0.6643	0.9177
	Self-generated	0.8622	0.8384	0.4226	0.7315	0.4309	0.6182	0.3110	0.6115	0.2881	0.7242	0.9177

Table 4: Comparison of zero-shot ICL performance on test sets between prompting with LOADS-selected label names versus the other three baseline approaches. We underline the highest model performance (statistically significant with paired chi-squared test).

the last decoder layer (averaged across the validation set).

3. Choose the label set with the lowest averaged kurtosis score.

The above LOADS-selected label set can then be used in the standard zero-shot ICL on test sets.

5.2 Evaluation

Setups We randomly sample 100 data points from the validation set for label selection and test the selected label sets on the official test sets.

- **Baselines:** We compare the model performance of using label words selected by LOADS to the following three approaches: (1) *Original label words*: we use the original label words from the dataset (i.e., the labels in Table 1), as it is the conventional and widely adopted practice; (2) *Original label words with a verbalizer*: after prompting the LLMs with the original label words, we employ our pool of candidate label words (See Section 3.1) as a verbalizer and incorporate their probabilities into the final probability of the same class; (3) *Self-generated label words*: we prompt the LLMs without providing any class options and select the candidate label words with the average highest probability at the first generated label token.
- **LLMs:** In addition to *Llama3* (8b), *Llama 3.1* (8b) and *Flan-T5-xl* (3b), we also examine whether LOADS could generalize to other model families and model sizes by including instruction-tuned *Gemma-2-it (9b)* and *Flan-T5-xxl* (13b).

Results Table 4 presents the zero-shot ICL model performances on stance classification and topic classification datasets with different label word selection strategies. The results demonstrate that employing LOADS to select label sets for zero-shot ICL prompts yields superior performance compared to other baseline approaches on most of the datasets. The improvement is consistent across NLP tasks and datasets, model architectures and sizes, as well as prompt templates[8].

Also, we observe limited benefits from adopting the verbalizer in post-processing, since LLMs tend to give the predicted label word tokens high probability in most of cases. Prompting with the self-generated label words rarely results in the best performance, while with the risk of leading to extremely low performance on certain datasets (e.g., perspectrum and snopes datasets with Llama 3).

Furthermore, potential data leakage could have significant impact on LOADS, indicated by the high performance achieved by the original label words on the AG News dataset with Flan-T5 models which was instruction-tuned with this dataset. It also aligns with the results in Table 3 where no statistically negative correlation was observed on the AG news dataset with Flan-T5-xl.

5.3 Analysis

Cross-lingual Transferability Previous research (Zhang et al., 2023a) has demonstrated that for non-English datasets, prompting with English task

[8] The analysis of prompt sensitivity can be found in the Appendix H.

instructions (including label words) while keeping the input in the original language often yields superior performance than non-English instructions. Therefore, we investigate whether the optimal English label sets selected by LOADS on English dataset can also enhance performance for non-English datasets in this scenario.

We manually translate the English rumoureval Twitter test set into French and Portuguese. We select the optimal label set based on LOADS with 100 randomly sampled data from the English rumoureval validation set. We then prompt the LLMs with English task instructions and French/Portuguese inputs. Table 5 presents the results, indicating that the optimal English label set identified by LOADS also effectively improves performance on non-English datasets when the instruction is provided in English.

Model	Method	French	Portuguese
Llama 3	LOADS	0.5020	0.4528
(8b)	*Original Label*	0.4544	0.4284
Llama 3.1	LOADS	0.4728	0.4278
(8b)	*Original Label*	0.3731	0.3679
Flan-T5-xl	LOADS	0.5137	0.3912
(3b)	*Original Label*	0.4189	0.3317

Table 5: Performance when LLMs are prompted with English instructions (including label options) and French/Portuguese inputs. Label options are selected by LOADS with English validation data.

Data Efficiency To show the merit of data efficiency of LOADS, we randomly sample 50, 100, 300, 500 or 1000 data points from the validation set and compare the rankings of the label sets based on LOADS. Due to the resource restriction, we conduct experiments on snopes (binary classification) and emergent (three-way classification) datasets with Flan-T5-xl and Llama 3.

The results show that the rankings of the top 5 label sets remain consistent across different sample sizes. It suggests that LOADS can achieve comparable performance even with a smaller number of unlabeled data samples than 100, further highlighting the data efficiency of our proposed method. We provide computational cost estimation of LOADS using 100 unlabelled samples in Appendix I.

Label Transferability We explore whether LOADS-selected label sets can be generalised across datasets or even models for NLP tasks such as stance classification where different label lexicons are used to represent the same classes across datasets. Specifically, for each dataset D_i on each LLM M_j, we select the optimal label set $L_{D_iM_j}$ through LOADS. To analyse whether the LOADS-selected label set on one dataset could be adapted to another related dataset with the same LLM M, we calculate the overlap of optimal labels ($L_{D_i\_M}$) between each dataset. Similarly, we examine the overlap of optimal labels ($L_{D\_M_i}$) between each model to explore whether the LOADS-selected optimal labels for dataset D could be adapted across LLMs. We only focus on the positive and negative stances to enable comparison across binary, three-way and four-way stance classification datasets.

Our results indicate that LOADS-selected label sets is transferable across datasets on the same LLM, highlighting the potential of leveraging LOADS to identify optimal label words with established related datasets, avoiding the need to collect samples for the target new dataset. For example, the positive-negative stance label pairs identified for Llama 3 is *endorse* and *deny* across all the stance classification datasets. However, we find that the label words selected for a specific dataset on one LLM often differ from those identified for another LLM, suggesting LOADS' dependency on the underlying model architectures and parameters.

In summary, the LOADS-selected label sets tend to be model-dependent rather than dataset-dependent. This observation aligns with the mechanism of LOADS, as the neurons and their distributions are inherently tied to the specific model. We hypothesize that this may suggest a correlation between the LOADS-selected label words and the LLMs' internal representation or understanding of the target NLP task or concept (e.g., what is stance), highlighting potential directions for future studies.

6 Conclusion

We study the impact of label options in the prompt for classification in zero-shot ICL, including lexical choice, label order, and label elaborations. We observe a significant effect of the lexicons used to represent label words in the prompt, also linking to the models' sensitivity to the label order. Through neuron activation analysis, we find that optimal label sets produce fewer outlier neurons in LLMs' feed-forward networks. We then propose LOADS, a novel method for selecting optimal label sets using activation distribution kurtosis. Prompting with LOADS-selected label sets consistently outperforms the use of original dataset labels across different models. Our approach is post-hoc, data-efficient and requires no gradient propagation or model fine-tuning. It also demonstrates cross-lingual transferability when using English instructions for non-English datasets. By showing that carefully selecting label sets based on neuron activation patterns can significantly enhance model performance without requiring additional training or labeled data, this paper has important implications for leveraging LLMs in zero-shot classification.

Limitations

Our experiments focused primarily on stance classification tasks. We chose this task because the label ambiguity is an identified challenge (i.e., label names could be replaced by a sufficient number of synonyms without altering their meanings and scopes in the original study), and it has sufficient datasets for empirical study. Although we have tested the generalisabity of our findings on topic classification, with more datasets released and new tasks proposed in future studies, studies could be conducted to explore whether our findings generalize to a broader range of classification tasks and domains. Also, although we examined multiple models from the Flan-T5 and Llama families, our study did not include other popular language models (such as Phi[9] or Mistral [10]) due to computational resource limitation. Expanding the range of models would provide a more comprehensive understanding of label option's impact across different architectures.

Another limitation of our study is English-language bias. Although we have explored the cross-lingual transferability to French and Portuguese on one dataset, more extensive multilingual testing is needed to ensure the approach's effectiveness across diverse languages and cultures.

Our method, while more efficient than gradient-based approaches, still requires running inference on a subset of data to compute activation statistics. This may be challenging for resource-constrained environments or very large models. The efficiency of our method may also be challenged when the classification task contains a very large number of class categories. Furthermore, we ensure the inclusion of samples for each class. The effectiveness of our method might vary when the validation set is highly imbalanced or even lack of data for the minority class. The effectiveness of different distribution metrics is out of scope but we acknowledge that it may have significant improvement for our method.

Lastly, while we focused on technical performance, future work should consider potential biases introduced by label set choices and their implications for fairness and inclusivity in classification tasks. Addressing these limitations in future research will help to further validate and refine our approach to optimal label set selection for zero-shot ICL.

Acknowledgments

This work was partially funded by EMIF managed by the Calouste Gulbenkian Foundation[11] under the "Supporting Research into Media, Disinformation and Information Literacy Across Europe" call (ExU – project number: 291191)[12] and the UK's innovation agency (InnovateUK) grant number 10039039 (approved under the Horizon Europe Programme as VIGILANT[13], EU grant agreement number 101073921). Yue Li was supported by a Sheffield–China Scholarships Council PhD Scholarship. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

References

Rachith Aiyappa, Shruthi Senthilmani, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2024. Benchmarking zero-shot stance detection with flant5-xxl: Insights from training data, prompting, and decoding strategies into its near-sota performance. *arXiv preprint arXiv:2403.00236*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. SuperTweetEval: A challenging, unified and heterogeneous benchmark for social media NLP research.

[9] https://huggingface.co/microsoft/phi-2

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

[12] exuproject.sites.sheffield.ac.uk

[13] https://www.vigilantproject.eu

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12590–12607, Singapore. Association for Computational Linguistics.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.

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

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland. Association for Computational Linguistics.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.

Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. 2024. Investigating language-specific calibration for pruning multilingual large language models. *arXiv preprint arXiv:2408.14398*.

Andrey Kuzmin, Markus Nagel, Mart Van Baalen, Arash Behboodi, and Tijmen Blankevoort. 2023. Pruning vs quantization: Which is better? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziqian Lin and Kangwook Lee. 2024. Dual operating modes of in-context learning. In *Forty-first International Conference on Machine Learning*.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Sheng Liu, Lei Xing, and James Zou. 2023. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022c. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024b. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Junyu Mao, Stuart E. Middleton, and Mahesan Niranjan. 2024. Do prompt positions really matter? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4102–4130, Mexico City, Mexico. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia. ELRA and ICCL.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Carolina Scarton, Diego Silva, and Kalina Bontcheva. 2020. Measuring what counts: The case of rumour stance classification. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 925–932, Suzhou, China. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, pages 1–13.

Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. 2023. Toward human readable prompt tuning: Kubrick's the shining is a good movie, and a good prompt too? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10994–11005, Singapore. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Alessandro Stolfo, Ben Peng Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima

Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2024. From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2341–2369, Mexico City, Mexico. Association for Computational Linguistics.

Zhichao Xu, Daniel Cohen, Bei Wang, and Vivek Srikumar. 2024. In-context example ordering guided by label distributions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2623–2640, Mexico City, Mexico. Association for Computational Linguistics.

Zhe Yang, Damai Dai, Peiyi Wang, and Zhifang Sui. 2023. Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13209–13221, Singapore. Association for Computational Linguistics.

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, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023a. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A Datasets

We summarise the datasets we used in our study in Table 6. For the stance classification datasets without official validation sets, we use the train/validation splits provided by Schiller et al., (2021). We utilize the official test set for AG news[14], and the official "train/test random split in the COLING 2022 paper" for TweetTopic dataset[15]. TweetTopic dataset has six class categories, potentially resulting in more than 4,000 different label sets if we consider only five synonymy words for each category (i.e., more than 12,000 experiments on three LLMs). Due to our limited computational resource, we experiment with three topics: *pop culture*, *daily life*, and *science & technology*.

Dataset Name	Source	# of Label Sets
scd	Debates	31
perspectrum	Debates	31
snopes	News	31
ibmcs	Debates + Wikipedia	31
vast	Debates + Artificial	62
emergent	News	62
semeval	Social Media	93
rumoureval	Social Media	248
arc	Debates	62
AG News	News	50
TweetTopic	Social Media	64

Table 6: Datasets and the number of label sets we experiment with for each dataset.

B Data Leakage

As far as we know, Llama 3 and Llama 3.1 are not supervised fine-tuned with any public stance classification datasets. Flan-T5 is fine-tuned on Super-NaturalInstructions dataset (Wang et al.,

[14] https://huggingface.co/datasets/sh0416/ag_news

[15] https://huggingface.co/datasets/cardiffnlp/tweet_topic_single

2022), containing two English stance classification tasks (Kobbe et al., 2020) (i.e., task 209 and 513). The tasks are formed as a binary (“in favor” and “against”) and a three-way (“in favor”, “against” and “neutral”) classification, respectively. There is no overlapping between these two datasets and our nine experimented datasets.

C Label Pool Creation

Stance Classification Following the pipeline we described in Section 3 of the main paper, we collect the seed label sets from the nine stance classification datasets (see Table 1). For the semeval dataset, the label set in the original paper ("favor, against, neither" in Table 1 in main paper) is slightly different from the set used in their published dataset ("favor, against, none"), so we consider both of them.

For positive and negative stance label names, we aim to acquire word-pairs with semantically opposite meaning. We first extract antonym for each positive and negative seed stance label from WordNet. Since we obtain limited antonyms in this way, Claude is then used to generate synonym for each seed positive-negative stance label pairs. An example of the prompt we used is: *Provide 5 different pairs of synonyms for "support" and "deny". They are supposed to be labels for stance classification.* We use WordNet to obtain synonyms for the rest of stance labels if there are any. For the label names that represent "neutral" stance in the original study, such as "observing" and "comment", we take "neutral" as their synonyms. Finally, we manually select the appropriate label names generated by Claude. The number of label sets we experiment with for each dataset is listed in Table 6.

Topic Classification Similarly, we follow the pipeline to collect and generate synonyms for each topic category. For TweetTopic dataset, since *pop culture* is a mixture of multiple sub-topics as discussed by Antypas et al.,(2022), we also consider the synonyms of the sub-topics. We use every possible combination of synonyms among topic categories for TweetTopic dataset. For AG news, there are total 160 combinations. We randomly sample 50 of them due to limited computational resources. The number of label sets we experiment with for two datasets is listed in Table 6.

D Decoding Strategies

We adjust the temperature (0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4) used for sampling-based decoding and compare their performances with the greedy search based performance for emergent and snopes dataset on Flan-T5-xl and Llama 3.

A temperature value larger than 1.0 – flattening the probability distribution – tends to harm the performance especially for Flan-T5-xl, which generates outputs irrelevant to stance. When temperature is lower than 1.0, introducing randomness in decoding through sampling may benefit the performance, but not significantly (in most of cases improvement is lower than 0.07). We summarise the maximum performance increase or decrease comparing with greedy search in Table 7.

Model Name	Temperature	snopes		emergent	
		+	-	+	-
Flan-T5	0.2	0.021	0.049	0.037	0.049
	0.4	0.055	0.066	0.066	0.068
	0.6	0.033	0.099	0.034	0.064
	0.8	0.033	0.145	0.066	0.107
	1.0	0.022	0.189	0.036	0.149
	1.2	0.037	0.264	0.046	0.196
	1.4	0.015	0.428	0.042	0.361
Llama 3	0.2	0.050	0.063	0.068	0.035
	0.4	0.044	0.047	0.073	0.073
	0.6	0.040	0.062	0.075	0.059
	0.8	0.027	0.062	0.066	0.107
	1.0	0.054	0.086	0.100	0.103
	1.2	0.038	0.072	0.069	0.171
	1.4	0.030	0.104	0.128	0.142

Table 7: The maximum performance increase (+) and decrease (-) if adopting sampling-based decoding rather than greedy search.

E Prompting with Label Explanation

We investigate whether the performance variance caused by different lexical choices of the label names could be mitigated or lowered by including the explanation of the label names in the prompt. We experiment with emergent and snopes datasets on Flan-T5-xl and Llama 3. We add the following class explanations in the prompt template after the class options for snopes and emergent datasets respectively: (1) snopes: *If the text supports that claim, answer with "{positive stance}"; if the text opposes the claim, answer with "{negative stance}"*; (2) emergent: *If the headline supports the claim, answer with "{positive stance}"; if the headline opposes the claim, answer with "{negative stance}"; if the claim is discussed in the headline but without assessment of its veracity, "{neutral stance}".*

We observe that including these label name explanations in the prompt may help with the label sets that achieve the lowest zero-shot performance. As for the snopes dataset, its worst performance would increase from 0.5400 to 0.555 (Llama 3, labels: *supportive* and *opposed*) or from 0.5766 to 0.6568 (Flan-T5-xl, labels: *for*, *against*). As for emergent, its lowest performance would increase significantly from 0.3877 to 0.5600 with Llama 3 (labels: *pro*, *con* and *neutral*). However, when using Flan-T5-xl, the inclusion of the class explanation even decrease the worst performance from 0.4870 to 0.3775 (labels: *support*, *deny* and *neutral*).

More importantly, the benefits from label explanations in the prompt would not close the gap between the optimal and sub-optimal label sets, comparing the above improved performance with the maximum performances in Table 2 in the main paper.

F Label Order Results

We present the averaged absolute performance difference after re-ordering the label names in the prompt in Table 8. The influence is limited on average.

Dataset	Flan-T5	Llama 3	Llama 3.1
perspectrum	0.0148	0.0372	0.0771
ibmcs	0.0195	0.0696	0.0961
snopes	0.0296	0.0772	0.1259
scd	0.0309	0.0288	0.0484
emergent	0.0211	0.0689	0.1578
semeval	0.0230	0.0304	0.0527
vast	0.0311	0.0465	0.0495
rumoureval	0.0355	0.0720	0.0439
arc	0.0152	0.0606	0.0612

Table 8: The average absolute performance change after re-ordering the label options in the prompt.

The maximum performance gain and drop on each dataset after re-ordering the label names for the top-k optimal and poor label sets with Llama3, Llama 3.1 and Flan-T5-xl are in Figure 2.

G Label Elaboration Results

We supplement the averaged absolute performance difference for each level of elaboration on Llama 3.1 and Flan-T5-xl in Table 9.

H Prompt Sensitivity Analysis of LOADS

To analyse the the prompt sensitivity of LOADS, we test it on different prompt templates, select the label set through LOADS, and then compare the performance with that on the label sets in the original dataset.

	Dataset	E_1		E_2		E_3	
		Opt.	Sub-opt.	Opt.	Sub-opt.	Opt.	Sub-opt.
Llama 3	perspectrum	0.016	0.018	0.010	0.015	0.017	0.009
	ibmcs	0.027	0.041	0.024	0.014	0.029	0.044
	snopes	0.055	0.040	0.054	0.026	0.029	0.018
	emergent	0.051	0.053	0.047	0.038	0.022	0.040
	rumoureval	0.095	0.036	0.084	0.020	0.058	0.102
	arc	0.017	0.024	0.015	0.021	0.035	0.049
Llama 3.1	perspectrum	0.027	0.020	0.048	0.025	0.037	0.022
	ibmcs	0.033	0.018	0.054	0.031	0.057	0.067
	snopes	0.015	0.021	0.032	0.020	0.034	0.033
	emergent	0.048	0.077	0.048	0.107	0.034	0.133
	rumoureval	0.041	0.040	0.037	0.032	0.039	0.035
	arc	0.032	0.046	0.029	0.036	0.042	0.046
Flan-T5-xl	perspectrum	0.009	0.026	0.010	0.021	0.013	0.021
	ibmcs	0.014	0.015	0.016	0.020	0.017	0.028
	snopes	0.020	0.026	0.022	0.032	0.033	0.016
	emergent	0.026	0.025	0.044	0.031	0.042	0.042
	rumoureval	0.059	0.040	0.060	0.031	0.086	0.018
	arc	0.034	0.042	0.025	0.038	0.031	0.041

Table 9: The average absolute performance change after elaborating for *optimal* or *poor* single-word label sets with Llama 3.1 and Flan-t5-xl (E_1, E_2, E_3 see Figure 1 in main paper).

Due to the computational resource constraints, we manually craft two prompts and test LOADS with the four binary stance classification datasets on Llama 3. In the two prompts, we replace *Given a [text1_name] and a [text2_name], detect the stance that the [text2_name] has towards the [text1_name]* (see Section 3.2 in main paper) with two different queries:

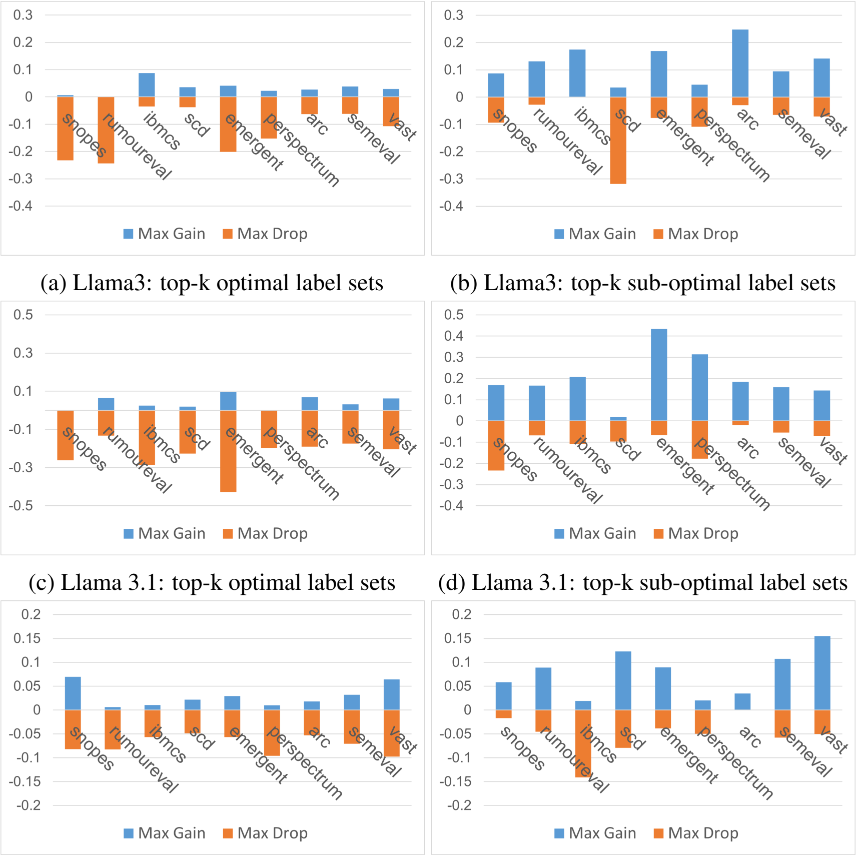
1. Prompt 1: *What is the stance of [text2_name] towards [text1_name]?*

2. Prompt 2: *What stance does [text2_name] take regarding [text1_name]?*

As shown in Table 10, although different prompts with the same label sets may result in performance changes as expected (compare with Table 5 in main paper), LOADS is robust to different prompts used for the label selection. The performance gap between LOADS-selected and original label sets tends to be similar across prompt templates.

I Computational Cost Estimation

Following previous work (Kaplan et al., 2020; Liu et al., 2022a), we estimate that a decoder-only LLM with N parameters uses $2N$ FLOPs per token for inference. We suppose that: (1) the input token length for the dataset we are interested in is L on average; (2) the number of candidate label sets is X; (3) 100 unlabelled texts are used for LOADS.



(a) Llama3: top-k optimal label sets

(b) Llama3: top-k sub-optimal label sets

(c) Llama 3.1: top-k optimal label sets

(d) Llama 3.1: top-k sub-optimal label sets

(e) Flan-T5-xl: top-k optimal label sets

(f) Flan-T5-xl: top-k sub-optimal label sets

Figure 2: The maximum performance gain (positive value) and drop (negative value) on each dataset after re-ordering the label names for the top-k optimal and sub-optimal label sets with Llama3, Llama 3.1 and Flan-T5-xl.

	Dataset	LOADS	Original Label
Prompt 1	snopes	0.5926	0.4984
	ibmcs	0.8737	0.7303
	perspectrum	0.8925	0.8658
	scd	0.6895	0.6860
Prompt 2	snopes	0.6191	0.5336
	ibmcs	0.8619	0.7523
	perspectrum	0.8921	0.8619
	scd	0.6836	0.6931

Table 10: Performance comparison on Llama 3 when using LOADS-selected label sets (*lowest kurtosis*) and using original label sets (*original label*) with prompt 1 or prompt 2. The higher performance is underlined.

Therefore, the total FLOPs taken by LOADS would be $2N * L * X * 100 = 200NLX$.

J Perplexity Analysis

As discussed in Section 2 in main paper, Gonen et al.,(2023) empirically show that zero-shot ICL performance is statistically negative correlated with the perplexity of the prompt with input. However, they did not take into account the label options in the prompt when calculating the perplexity. Therefore, we further investigate whether the perplexity is also correlated with the variance zero-shot ICL performance caused by different label names.

Specifically, we use the prompt template in Section 3.2 in main paper, and calculate the perplexity of prompts with inputs and different label sets. Following Gonen et al.,(2023), for each label set, we average the perplexity over the dataset. And then we adopt spearman correlation test between the averaged perplexity scores and model performances. Due to the computational restriction, we experiment with all the binary datasets on Flan-T5-xl and Llama3-8b. Since Flan-T5 is an encoder-decoder model where perplexity has a loose definition, we treat the encoder input as an empty string when calculating perplexity.

The results in Table 11 indicate that there is no statistically significant correlation between prompt perplexity and model performance if considering different label sets in the prompt.

		perspectrum	ibmcs	snopes	scd
Llama 3	*coefficient*	0.0068	0.1641	0.0394	0.1698
	p value	0.9707	0.3774	0.8303	0.3608
Flan-T5	*coefficient*	0.0738	0.1733	-0.0500	-0.2273
	p value	0.6929	0.3511	0.7892	0.2187

Table 11: Spearman correlation between model performance and prompt perplexity. P-values are all larger than 0.05, indicating no statistical significance.

K Label Attention Key Similarity Analysis

In this section, we explore whether the closely related observation on few-shot ICL could be directly adopted to zero-shot ICL. Specifically, we focus on the study discussed in Section 2 in main paper, where Wang et al.,(2023) suggest that the when the LLM is prompted by demonstration with examples in a few-shot ICL setting, the model is likely to confuse the label categories if their key vectors in the attention modules are similar to each other.

Since this finding is easier to be tested on binary datasets, we experiment with the binary datasets on Llama 3 and Flan-T5-xl. We extract the key vectors in the attention module in each layer for each label name in the prompt. Then we calculate the cosine similarity between the vectors of two label names. Finally, we use spearman correlation test between similarity scores and model performances. As shown in Table 12, we do not observe statistically significant correlation between model performance and label key vector similarities.

		perspectrum	ibmcs	snopes	scd
Llama 3	*coefficient*	-0.0181	-0.0051	-0.2791	-0.1696
	p value	0.9244	0.9784	0.1283	0.3702
Flan-T5-xl	*coefficient*	-0.0595	0.0223	-0.0209	-0.0992
	p value	0.7503	0.9047	0.9108	0.5953

Table 12: Spearman correlation between model performance and label's key vector similarity.

L Layer-wise Output Projections Analysis

We hypothesize that LLM may jump to the output prediction at last layers when a sub-optimal label set is used in the prompt. Therefore, we extract the hidden states from each decoder layer and project them on the model vocabulary, so that we obtain the ranked position of the final predicted label's token in each layer (Elhage et al., 2021; Geva et al., 2022).

we observe that the hypothesis indeed holds in certain cases. We show an example on rumoureval dataset where we compare the averaged rank of the correctly predicted label *comment/neutral* in each decoder layer when Flan-T5-xl is prompt to choose from *support, deny, query, comment* or *support, deny, query, neutral*. Label set *support, deny, query, comment* performs worse than the set *endorse, deny, query, neutral* on this dataset. As shown in Figure 3, when using the relatively optimal label set *endorse, deny, query, neutral*, the rank of the final predicted label tends to move closer to the top at an earlier stage.

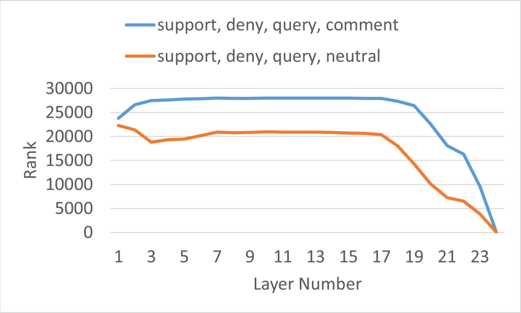


Figure 3: The rank of the final correctly predicted label (*comment* or*neutral*) when Flan-t5-xl is prompted with two different label sets for rumoureval dataset.

M Human Translation Details

To translate the English Twitter rumoureval test set into French and Portuguese, we recruit volunteer students from translation studies in Brazilian and French universities. The students are given gift vouchers (0.6 pounds per tweet). Consent has been obtained from the students and our study has received approval from the Ethics Committee of our university.

We instruct the students to translate the tweets accurately, and preserve the original meaning, context, and tone of the tweet. They are also encouraged to leave notes for their translations. The translations are finished on Google Sheets.