Token-based Decision Criteria Are Suboptimal in In-context Learning

Anonymous ACL submission

Abstract

In-**C**ontext **L**earning (ICL) typically utilizes classification criteria from output probabilities of manually selected label tokens. However, we argue that such token-based classification criteria lead to suboptimal decision boundaries, despite delicate calibrations through translation and constrained rotation applied. To address this problem, we propose Hidden Calibration, which renounces token probabilities and uses the nearest centroid classifier on the LM's last hidden states. In detail, we assign the label of the nearest centroid previously estimated from a calibration set to the test sample as the predicted label. Our experiments on 6 models and 10 classification datasets indicate that Hidden Calibration consistently outperforms current token-based baselines by about 20%∼50%, achieving a strong state-of-the-art in ICL. Our further analysis demonstrates that Hidden Calibration finds better classification criteria with less inter-class overlap, and LMs provide linearly separable intra-class clusters with the help of demonstrations, which supports Hidden Calibration and gives new insights into the principle of ICL.

1 Introduction

In-**c**ontext **L**earning (ICL) (Dong et al., 2022) is a few-shot learning paradigm without any model parameter updates on **L**anguage **M**odels (LMs). In detail, as shown in Fig. 1-A, B, given a prompt consisting of demonstrations and a query, LMs conduct causal language modeling operation from the prompt to assign probabilities to the label token candidates designed by hand, and ICL chooses the one with the highest probability as the prediction.

One well-known issue of ICL is that the predicted probabilities are biased (*under-calibrated*), leading to a decrease in prediction performance (Fei et al., 2023; Han et al., 2022; Zhao et al., 2021; Zhou et al., 2023). To solve this issue, previous work *calibrates* the predicted label

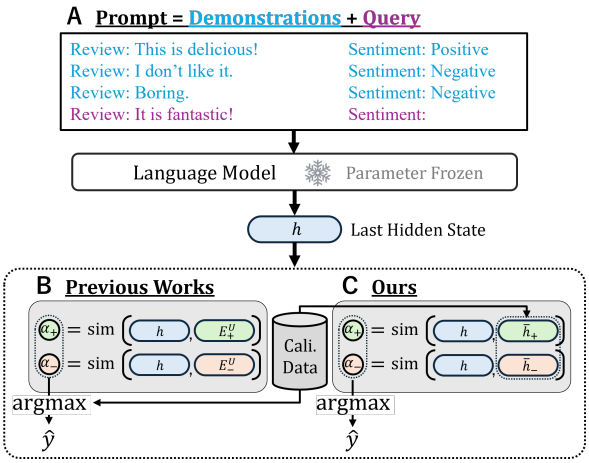

Figure 1: In an ICL diagram, **A.** The prompt of ICL consists of a combination of demonstrations and a query. LMs encode the prompt into the last hidden state h, then **B.** Previous works use the un-embedding vectors of the label tokens (E_+^U and E_-^U) to decode the h to prediction $\hat{y}$, then calibrations are used to adjust the predicted logits. **C.** Our work uses the calibration dataset to calculate centroids ($\bar{h}_+$ and $\bar{h}_-$) to decode the h.

token probabilities by performing affine transformations with estimated parameters to adjust these probabilities for more precise predictions.

These previous works and also the vanilla ICL are based on a potential assumption: the affine manifolds spanned by the decoding vectors in the LM head (*un-embedding vectors*) of *manually selected* label tokens are good subspaces of the hidden space to distinguish hidden vectors with various label appropriately, so that the label token probabilities decoded from these subspaces are good classification logits. However, some practices have pointed out that randomly changing label spaces doesn't critically influence ICL performance (Min et al., 2022c; Wei et al., 2023), which means the selected label subspaces are trivial and arbitrary. Although using the label un-embedding with natural task semantics (e.g. "positive" and "negative") seems intuitive,

it should be noted that we have no reason to believe that these label tokens have any guarantee for generating classification logits with good decision boundaries, even if various delicate calibrations are used to move these boundaries *inside the subspaces*. This makes us suspect that: **Utilizing manually selected label probabilities as classification criteria may not be good ICL practices**.

Previous work has shown that using the output probabilities of the *full vocabulary* increases ICL performance (Xu et al., 2022; Abbas et al., 2024). This is a good start to avoid the manually selected mapping subspace, but there is still doubt that output probability distributions are not informative enough for classification. Therefore, we utilize the last hidden states instead, which are informative precursors of the token probabilities.

In practice, we propose Hidden Calibration, training centroid classifiers on the last hidden states of ICL prompts. As shown in Fig. 3, **during the training**, we build standard ICL prompts similarly to Fig. 1-A from a calibration set and input them into an LM to get the last hidden states of the last tokens of ICL prompts. Then, we calculate the centroids of the last hidden states w.r.t. the queries' label to get a centroid for each label, as an anchor for inference. **During the inference**, we input the test prompt, find the nearest centroid to the last hidden states of the test prompt, and assign the corresponding label of the centroid as the prediction.

Empirically, Hidden Calibration improves the ICL performance by approximately more than 20% on 10 text classification datasets and 6 modern LMs, with an equal computational cost with previous calibration methods. To the best of the author's knowledge, Hidden Calibration consistently outperforms the calibration baselines, achieving a strong state-of-the-art in ICL. Additional experiments indicate that Hidden Calibration effectively alleviates the stress of prompt engineering, performing robust accuracy under different prompt designs.

Moreover, our subsequent analysis indicates that Hidden Calibration does find better logits-mapping subspaces that effectively separate data points. In detail, we find that the distribution of classification logits calculated from Hidden Calibration have less inter-class overlapping than from label probabilities, while such overlapping is proportional to the lower bound of the classification error. This suggests Hidden Calibration finds subspaces with essentially better classification performance.

Furthermore, we investigate the principle of Hidden Calibration, that is, the reason why a simple centroid-based linear decision boundary can divide the ICL hidden state properly. We find that LMs provide linearly separable clusters in the hidden states w.r.t. query labels, while demonstration can promote such a process.

Our contributions can be summarized as:

- We analyze the previous calibration practices on ICL, and find their consistent limitations: Using manually selected labels as the projecting subspaces for classification criteria is often under-guaranteed.
- We propose Hidden Calibration to address such a problem, eliminating the unreliable decoding on the handmade label, and using the centroid classifier on the LM's last hidden states. Our experiments suggest that Hidden Calibration reaches a new state-of-the-art.
- Our further analysis indicates that Hidden Calibration does find better classification criteria with less inter-class overlap, and LMs provide linearly separable intra-class clusters with the help of demonstrations, which supports Hidden Calibration.

2 Background

This section reviews previous work on ICL and denotes their mathematical descriptions as an introduction to our main motivation.

2.1 In-context Learning

Prompting. Given a few-shot natural language classification dataset (*demonstration set*) $\mathcal{D} = \left\{\left(x^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y}\right\}_{i=1}^{n}$, where $x^{(i)}$ and $y^{(i)}$ are the input sequence and label token of i-th data point, and $\mathcal{X}, \mathcal{Y}$ is the input and label space, respectively, we sample a set of k samples $\mathcal{D}^{de} = \left\{\left(x^{(c_i)}, y^{(c_i)}\right)\right\}_{i=1}^{k}$ from $\mathcal{D}$ with an index set $\{c_i\}_{i=1}^{k}$ for a given query x^q. Then, we use a template T to concatenate them in a natural language form into a prompt token sequence: $s = T\left(\mathcal{D}^{de}, x^q\right)$, as shown in Fig. 1-A.

Encoding. A decoder-structured LM receives the prompt token sequence s and encodes it into the *last* (from the last Transformer layer) hidden state matrix as $H \in \mathbb{R}^{|s| \times d}$ with a length of token $|s|$ and embedding dimension of d. We denote the hidden state of the last token as $h = H_{|s|} \in \mathbb{R}^{d}$.

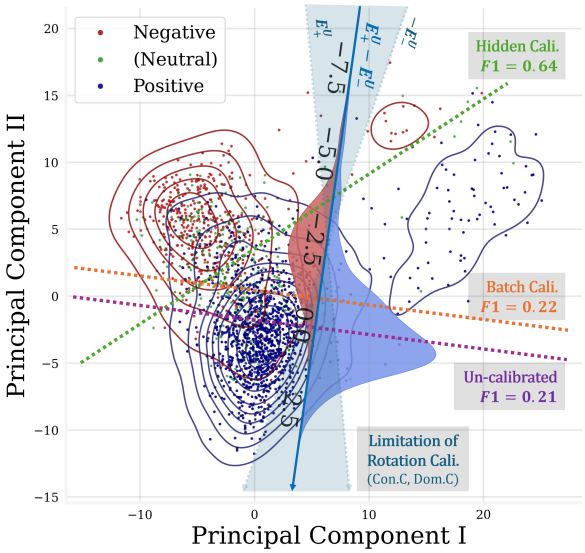

Figure 2: Token probability-based decision boundaries (original & batch calibrated) are suboptimal comparing to centroid-based boundary. Points and contour lines are ICL's last hidden states and kernel densities mapped by Principal Component Analysis. Oblique coordinate axis is the direction of the un-embedding difference of label tokens $\left(E_+^U - E_-^U\right)$, where the kernel densities of mapped data points are plotted. The rotating calibration by $A \neq \mathbf{1}$ (e.g. Contextual Calibration, Domain Calibration) has a limited feasible mapping direction[3].

Constrained Decoding. In a typical ICL setup, one chooses the un-embedding vectors of the label candidates in the output head[1] as the decoding subspace, i.e., For each label l, the similarity $\alpha_l = \mathrm{sim}(h, E_l^U)$ (usually the dot-product similarity) between h and each un-embedding vector E_l^U is calculated as the output logits α_l, as shown in Fig. 1-B for a binary classification example. Then, the label with the highest logits is chosen as the prediction $\hat{y}$, that is: $\hat{y} = \underset{l \in \mathcal{Y}}{\mathrm{argmax}}\, \mathrm{sim}(h, E_l^U)$.

2.2 Token-probability Calibration for ICL

However, Zhao et al. (2021) find that simply using the original logits for classification can not lead to a good ICL practice, since these logits have considerable prior bias and often tend towards specific labels even if the query is blank or meaningless (Zhao et al., 2021; Fei et al., 2023). Some calibrations have been proposed to mitigate such bias in a linear form: first, the logits are transformed into probabilities as $p = \mathrm{softmax}([\alpha_1, \alpha_2, \ldots, \alpha_{|\mathcal{Y}|}])$, then affine-transformed as calibrated classification criteria $p' = A \odot p + B$, where $A, B \in \mathbb{R}^{|\mathcal{Y}|}$ is the calibration terms estimated from m training examples from a calibration set, and $\odot$ is the Hadamard multiplication. Various estimations for A and B are used: Some practices use examples with pseudo queries terms (Fei et al., 2023; Zhao et al., 2021), while other practices use Gaussian estimation on real prompts (Han et al., 2022) or the mean value of p during the inference (Zhou et al., 2023).

However, such calibrations are affine transformations on label token probability, without modifying the E_l^U, allowing only very limited improvement towards ICL inference (to be discussed in §3.1).

3 Methodology

Based on the above background, in this section, we demonstrate the limitations of the above calibrations, and then propose Hidden Calibration to address such limitations fundamentally.

3.1 Token Probabilities Are Not Good Classification Criteria

To better understand the limitations of the token probability-based paradigm, we show a prototypical visualization of the hidden states of ICL prompts. Specifically, we input 2,048 ICL prompts (with $k = 8$) built from of `SemEval 2014-Task 4 Restaurant` (Pontiki et al., 2014) into OPT-2.7B (Zhang et al., 2022) and plot the last hidden states of the last token on a 2D-Principal Component plane in Fig. 2 (Appendix B.4).

As a simple 2-way case, focusing on the data points labeled "positive" and "negative", we plot the difference direction $\left(E_+^U - E_-^U\right)$ between the un-embedding vectors of these two label tokens[2]. Then, the coordinates of the projected hidden states in this direction are the difference of predicted logits between these two labels, serving as a *token-based classification criteria*, i.e., when the difference is positive, a "positive" label will be assigned, and vise versa. Therefore, in this visualized scenario, the original decision boundary is the orthogonal line at the zero point, the batch calibrated boundary (Zhou et al., 2023) is always parallel to it, and the other calibrations (Contextual Calibration (Zhao et al., 2021), Domain Calibration (Fei et al., 2023)) produce rotated mapping directions

[1]We omit the bias term in the output head (if any) for the sake of simplicity. It can be overridden by a fixed-to-one dimension, or covered by the calibration described below.

[2]Notice that Principal Component Analysis is an orthogonal transformation, keeping the dot-product and normal line fixed (In fact, beyond orthogonal transformations, they are also centralized. Therefore, the projection axis does not necessarily pass through the coordinate origin). See Appendix B.4.

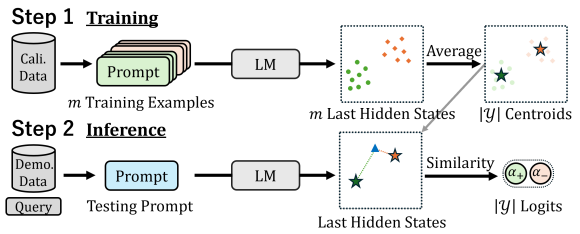

Figure 3: The diagram of Hidden Calibration. **Step 1:** Calculating the hidden state centroid of each label. **Step 2:** Find the label of the nearest centroid of the text sample to be the prediction.

$(A_+E_+^U - A_-E_-^U)$, by term A and thus rotated decision boundaries, where $A_+, A_- > 0$ limits[3] the direction inside the closure of E_+^U and $-E_-^U$.

Intuitively, as shown in Fig. 2, the token-based decision boundaries cannot effectively classify these data points, which is due to the inherent direction of the token un-embedding vectors, regardless of calibration applying limited affine transformation. A straightforward better linear boundary is the equidistant points between both classes' centroids, plotted as the green line.

3.2 Hidden Calibration

Motivated by the visualization, we propose Hidden Calibration, using the centroid similarity as the classification logits. In practice, we use a 2-step paradigm as shown in Fig. 3, first, as training, we calculate the centroid of the last hidden states of data points within each class on some prompt-label training examples. Then, in the inference, we select the closest centroid of the test prompt's hidden state as the prediction.

In detail, **(1) Training:** Given a calibration set with m supervised prompt-label pair $\{(s^{(i)}, y^{(i)})\}_{i=1}^m$, where the $s^{(i)}$s (*training examples*) are standard ICL prompts with k demonstrations, and $y^{(i)}$s are the ground-truth labels of corresponding $s^{(i)}$s' query. We input each training example to LM, and extract the last hidden state $h^{(i)}$. Repeating on the whole training example set, we can get a supervised hidden state set $\mathcal{H} = \{(h^{(i)}, y^{(i)})\}_{i=1}^m$. Then, we calculate the centroids of label l as: $\bar{h}_l = \mathbb{E}_{(h^{(i)}, y^{(i)}) \in \mathcal{H}, y^{(i)}=l}\left[h^{(i)}\right]$.

Then, we utilize the calculated centroids in **(2) Inference:** Given a test ICL prompt, we input it into the LM and get the last hidden state h, we calculate the similarity between h and every centroid $\bar{h}_l$ as the centroid-based logits α_l. In practice, the additive inverse of Euclidean distance is used as the similarity (that is, $\alpha_l = -\left\|h - \bar{h}_l\right\|_2^{\frac{1}{2}}$), while Appendix D.1 shows that Hidden Calibration acts equally on cosine similarity measures. We assign the label with the highest logits as the prediction.

Notice that another intuitive solution to the problem in §3.1 is utilizing the logits or probabilities of the *whole vocabulary*, as shown in previous works (Xu et al., 2022; Abbas et al., 2024). However, since the input, hidden states, and logits form a Markov chain, no input-relevant information gain is conveyed to the full-vocabulary logits. Moreover, the dimensionality of the full-vocabulary logits is typically significantly greater than the hidden states, therefore we choose the hidden states, a dense and informative precursor of token probabilities, as the classification feature.

Moreover, more complex classifiers, such as a KNN classifier, or a perceptron, can be used on the last hidden states instead of a centroid classifier. However, we choose the centroid classifier as the simplest implementation to avoid attribution confusion and enable the following analysis. Also, a centroid classifier has a minimal computation cost to fit the scenario of low-resource (Appendix D.2).

4 Experiments & Main Results

In this section, we prove the validity of Hidden Calibration by testing its classification performance on 6 models and 10 datasets. Hidden Calibration outperforms all the baselines and reaches a strong state-of-the-art of ICL with high efficiency in calculation, data, and prompt engineering.

4.1 Experimental Settings

Models. We use 6 models: OPT-2.7B (Zhang et al., 2022), LLaMa 2 (Touvron et al., 2023) (6.9B, 13B, 34B), LLaMa 3 (AI@Meta, 2024) (8B) and GPT2-XL (Radford et al., 2019) (1.6B). Models larger than 10B are quantized.

Baselines. We use 6 baselines from the previous works, with 4 label token-based methods (None, Con.C, Bat.C, and Dom.C) and 2 whole vocabulary probabilities-based methods (KNN and Cent.C). Details can be found in Appendix B.2.

Datasets. We use 10 commonly used classification datasets with some of the overlength data

[3] In current practices, the A are calculated from reciprocals of probabilities, which are positive-definite (Note that the calibration is trivial when A is not positively definite: the label with negative A components will never be assigned), and usually do not have significant relative values.

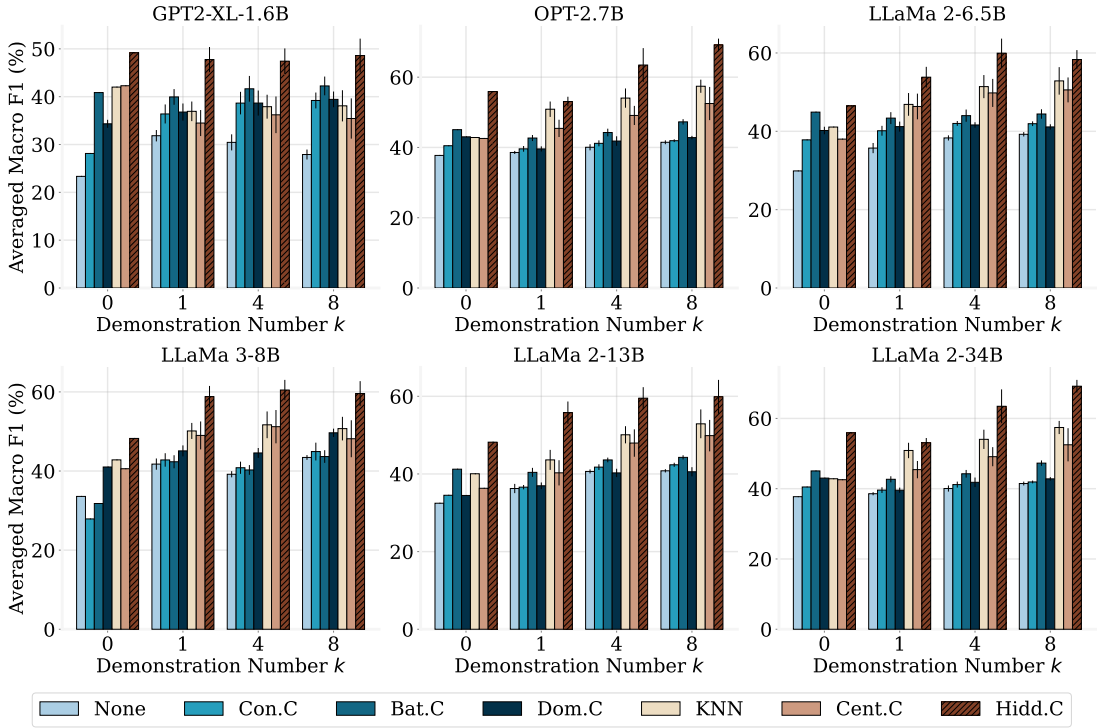

Figure 4: The classification performance (Macro F1(%)) of 6 models averaged on 10 datasets. Hidden Calibration (Hidd.C) is a new state-of-the-art of ICL, where demonstrations consistently improve the performance.

points excluded. See Appendix B.1 for details.

Other details. All the model checkpoints and datasets are loaded from `HuggingFace`. Macro F1 is used as the classification metric. We use a simple template to generate the prompt, see Appendix B.3. We set $m = 16|\mathcal{Y}|$ training examples (16 examples per class), and for fairness, every baseline is given equal training examples for calibration. All the experiments are repeated 5 times.

4.2 Main Results: Hidden Calibration is A New State-of-the-art of ICL

The tested classification performance of Hidden Calibration and baselines is shown in Fig. 4, where Hidden Calibration (Hidd.C) consistently outperforms all the label token-based or vocabulary-based baselines. Comparing to the vanilla ICL (None), Hidden Calibration produces an improvement up to around 100%. In general, compared to the highest baseline, Hidden Calibration improves the performance by approximately 20%. Detailed numeric and Accuracy results are in Appendix. C.1.

Especially, compared to the Cent.C baseline proposed by us for a controlled trial, which conducts the same calculation but uses the whole output token probabilities instead of the hidden states, Hidden Calibration outperforms, which confirms our idea that token probability distribution is a less informative classification feature mentioned in §3.2.

4.3 Efficiency: Low Complexity towards Time, Space, Data, and Prompting

Time and Space Complexity. Intuitively, Hidden Calibration has little additional computational cost compared to the calibration baselines, since they require almost equivalent feedforward calculations, making it competitively efficient as calculated in Appendix D.2.

Training Data Complexity. Hidden Calibration requires additional annotated data compared to label token-based calibration methods of the same scale in an acceptable range. In detail, in label token-based calibrations of k demonstrations, k supervised data is needed with a synthetic query to build a training example, while Hidden Calibration needs a real query **for each label**, requiring one more supervised data. However, for classification tasks, preparing an example for each class can be easily done whether in an industry or laboratory scenario, furthermore, these data and trained centroids can be reused (Appendix D.3) to further reduce the requirement of annotated data.

Data Efficiency. Regarding data efficiency of training examples, we repeat the experiments with

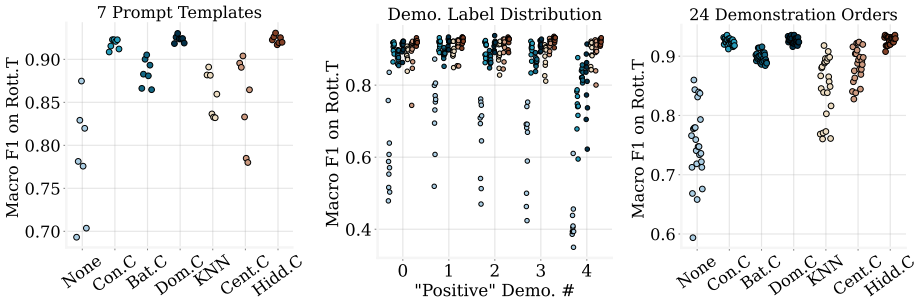

Figure 5: Sensitivities on (**left**) prompt template, (**middle**) demonstration label distribution, and (**right**) demonstration order (measured by performance distribution) on LLaMa 2-6.9B and `Rotten_Tomatoes`. Legend is consistent with Fig. 4, omitted.

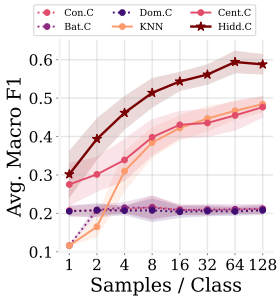

Figure 6: Classification performance against training examples.

various m on OPT-2.7B (see Appendix B.7 for details), from 1 to 128 calibration examples per class. The results are shown in Fig. 6, which indicate that Hidden Calibration stably benefits from the size of the calibration set, while even one sample per class can still make it outperform. Meanwhile, token-based methods can not benefit from scaling the calibration dataset, which also makes Hidden Calibration a better practice when a considerable amount of supervised data can be accessed: data can be used to construct the centroid to improve the classification in a linear cost, rather than increase the demonstrations in a quadratic cost.

Prompting Complexity. Hidden Calibration reduces the pressure on prompt engineering. We investigate this for Hidden Calibration and baselines on LLaMa 2-6.9B and `Rotten_Tomatoes` ($k = 4$) in three aspects: **(1) Prompt template** (Voronov et al., 2024). We select 7 different prompt templates (shown in Appendix B.3) and test ICL performance on them, shown in Fig. 5 (left). **(2) Label distribution in demonstrations.** We construct prompts with various numbers of "positive" demonstrations presented, and test ICL performance shown in Fig. 5 (middle), **(3) Demonstration order** (Lu et al., 2022). We enumerate the full arrangements from a fixed demonstration set of $k = 4$, and test the ICL performance using each demonstration arrangement, shown in Fig. 5 (right). All the results show that: compared to baselines, Hidden Calibration keeps narrow and high-performance distribution against all the three variables, i.e., Hidden Calibration stably works for various prompts, providing higher efficiency on prompt designing.

5 Analysis

This section attempts to enhance our understanding of Hidden Calibration through comprehensive observations. **(1)** Similar to Fig. 2, we measure the inter-class overlapping area on data points projected into classification criteria, to find whether Hidden Calibration maps data into logits with smaller inter-class overlap, i.e., better separability. **(2)** We further investigate why simple linear boundaries can effectively classify ICL hidden states, as happens in Hidden Calibration. We find that LMs provide a primary linear clustering in hidden states responding to query classes, and such clustering is enhanced by more demonstrations.

5.1 Effectiveness: Hidden Calibration Finds Criteria with Lower Overlap

In Fig. 2, we projected all the "positive" and "negative" labeled data points into the difference of the label logits (*vanilla classification criteria*) on the oblique extra coordinate axis, then a significant *overlap* between the projected data point cloud in two classes can be observed, making it difficult to find suitable classification boundaries vertical to the projection direction. In this section, we quantify the intuitive observation into the area of such overlap as a metric of classification criteria.

We first decompose the multi-classification dataset into all possible binary classification combinations w.r.t. the ground-truth labels. Then, for each combination, we sample standard ICL prompts with queries labeled with a specific one of the selected two labels, and map the prompts into the difference of classification **probabilities** across these two labels by the calibration method to be investigated. To get a continuous distribution of the difference, we run kernel density estimations on the calculated difference. We repeat this processing on the other label in the binary combination, and get two density estimations in total for both labels in a binary combination, as shown in Fig. 7 (curves). Then, we calculate the overlap area of

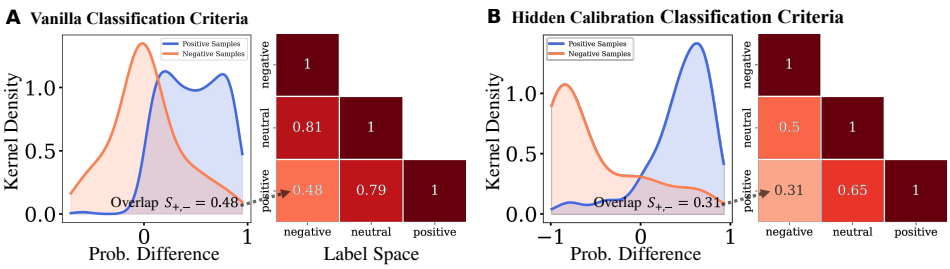

Figure 7: Demonstration of the overlap calculation with GPT2-XL on `SemEval 2014-Task 4 Restaurant`, $k = 4$. **Curves:** The kernel density of probability difference of l_1 ="positive" and l_2 ="negative". **Heatmaps:** The overlap of 2-combinations (we plot the combination with the same label with overlap 1, but omit them in averaging).

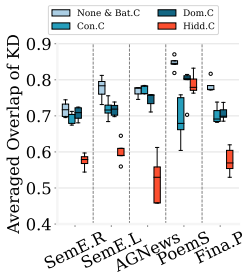

Figure 8: The Averaged Overlap of both criteria on GPT2-XL and 5 datasets.

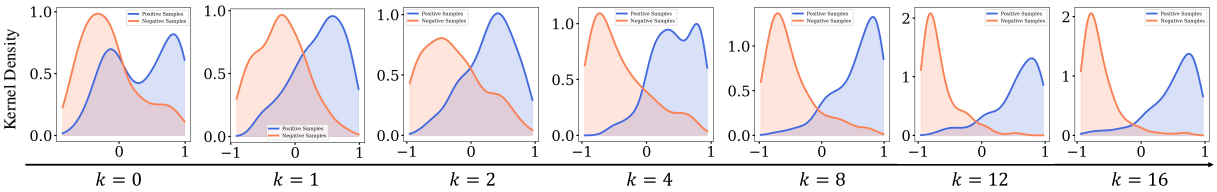

Figure 9: In-context learning dynamics visualized by overlap on OPT-2.7B and `SemEval 2014-Task 4 Laptops` (Pontiki et al., 2014). The data points appear in clusters responding to their query labels when no demonstrations are given and gradually converge to the centroid w.r.t. the demonstrations number.

these two kernel density curves. We repeat such processing for each binary combination as shown in Fig. 7 (heatmaps), and the final Averaged Overlap is the macro average of overlap area among all possible binary combinations (operation details are in Appendix B.5).

The overlap area of the two distribution curves is double to the *lower bound* of the classifier's error rate among these two labels (Appendix B.5.3), so Averaged Overlap is an intuitive metric of the classification criteria: The larger the overlap, the more difficult it is for the classifier, even (further) calibrated or ideal, to classify data points correctly, resulting in a potential decrease in accuracy.

We measure the Averaged Overlap of 4 un-quantized models on 5 datasets (see Appendix B.5.2 for experimental details). The result on GPT2-XL is shown in Fig. 8 (see Appendix C.2 for other models), where the Averaged Overlaps from token-based methods are consistently high, causing that better classification performance cannot be achieved on such criteria, which confirms our hypothesis in §3.1. Meanwhile, the overlaps from Hidden Calibration is much less than from token-based methods, meaning that Hidden Calibration produces better classification criteria with better possible classification performance than the token-based methods, even if delicate calibrations transfer or rotate these classification boundaries.

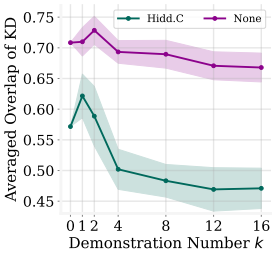

Figure 10: The Averaged Overlap on OPT-2.7B and 5 datasets against the demonstrations number.

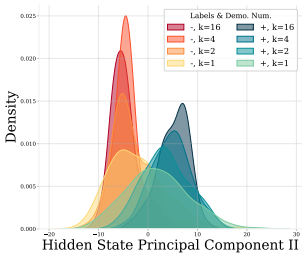

Figure 11: Hidden state convergence w.r.t. k of Fig. 9 visualized on the direction of principal component II.

5.2 Principle: The Inner Linear-separability

In the practice of Hidden Calibration, simple linear boundaries are used to classify ICL examples, raising doubt on the linear separability of hidden states. In this section, We find that LMs primarily produce linearly separable hidden state clusters corresponding to the ground-truth label, and the demonstrations facilitate this process.

As an intuitive visualization, we plot curves same as the Fig. 7 but with various number of demonstrations k to visualize the *in-context learning dynamics* in Fig. 9, where we find that: **(1)** the data points have a little linear separability when $k = 0$, and **(2)** such linear separability is being enhanced following the increment of k, performing intra-class converging dynamics.

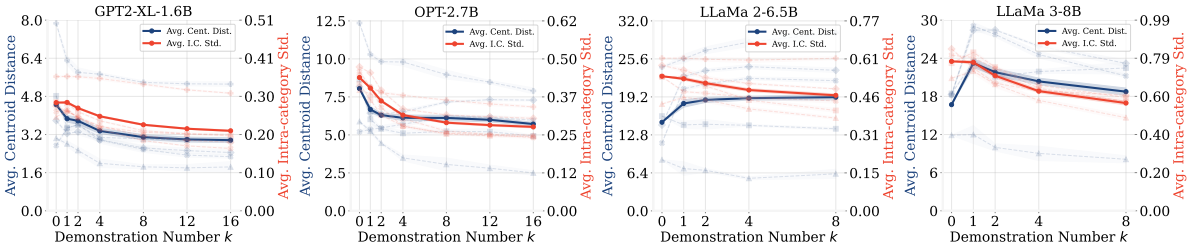

Figure 12: The averaged intra-class standard error of data points and the inter-class averaged centroid distance against k. **Solid lines**: means on 5 datasets; **Dashed lines**: Individual results for each dataset.

We further characterize this process. First, we calculate the Averaged Overlap similar to §5.1 against k in Fig. 10. We find that the token-based overlaps remain high and stable w.r.t. k, which indicates that the token-based methods can not benefit much from the demonstrations. However, the overlaps from Hidden Calibration significantly decrease with the increase of k, indicating that Hidden Calibration benefits from the demonstrations as expected, aligning with our observations in §4.2.

More generally, we visualize the last hidden states from similar inputs of Fig. 9 on the **second** principal components of hidden states to get an essential observation in Fig. 11, where as k increases, the hidden state shows clear intra-class cohesive dynamics, enabling linear classification through the clustering of hidden states.

More directly, on the last hidden states, we measure the intra-class standard error and the inter-class averaged centroid distance against k (see Appendix B.6 for details), as a joint measurement of intra-class aggregation and inter-class aggregation. The results are shown in Fig. 12, where the two curves are both diminishing, showing an obvious intra- and inter-class aggregation trend w.r.t. k. However, the inter-class aggregation has weaker and less persistent decreasing trends, presenting only in early demonstrations, or even ascending, which indicates that demonstration enhances intra-class clustering stronger than the inter-class aggregation, which is beneficial to linear classification. Moreover, a model with more parameters shows a stronger difference between these aggregations.

6 Discussion

Conclusion. In summary, we analyze the current token-based ICL decision boundaries and point out a limitation of using token probability to decode ICL prediction. To address such a drawback, We propose Hidden Calibration by decoding the classification logits from centroid classifiers on LM's last hidden states. Our experiments show that Hidden Calibration is a new state-of-the-art of ICL, with high efficiency on time & space, data, and prompting. Then, we confirm that Hidden Calibration indeed creates better classification logits by reducing the inter-class overlap. Moreover, we discover the hidden state convergence promoted by demonstrations, as an explanation of the principle of the performance improvement by a single linear classification boundary in Hidden Calibration. We hope this work can inspire exploration of the ICL by investigating the hidden state instead of token probabilities, and update the community's understanding of ICL calibration.

Comparison to Previous Works. (1) Comparison to Probe Methods. One concern is that our work can be regarded as a degraded linear probe (Abbas et al., 2024) of the hidden states. However, we believe our work has more advantages: In terms of application, we use fewer samples and require no gradient-based training, which makes our method more user-friendly, efficient, elegant, and interpretable. In terms of theory, compared to fitting a universal approximation (Hornik et al., 1989), our method and settings fully utilize the hidden state convergence on decoder LMs (described in §5.2), making it a true ICL practice. **(2) Comparison to Supervised Fine-tuning.** Some practices (Gu et al., 2023; Min et al., 2022b) build training objectives to fine-tune models for better ICL performance. These efforts are efficient but bulky, while our work avoids such an enormous overhead, making it more usable and elegant. **(3) Comparison to Other Calibrations.** Our method is a disruptive innovation for methods based on token probability (even the ones based on the whole vocabulary). Experimental comparisons of these methods have been given in §4.

For more related works, refer to Appendix A.

7 Limitations

Due to computability limitations, we are not able to compare the performance of Hidden Calibration with the baseline based on supervised fine-tuning. However, we believe that Hidden Calibration is not within the same methodology as the fine-tuning method, due to the significant difference in computational cost. So such a lack of comparison will not seriously hurt the soundness of this paper.

We prove that human intuition in the label token choice is not reliable. However, we have not eliminated such human intuition completely from the ICL loop: when we build prompts, we still choose the label token. How to automatically select the optimal label token in the prompt will be an important issue, leaving as future research directions for improving the performance of ICL further.

Other label probability calibrations (e.g. Batch Calibration) can be combined with Hidden Calibration for further performance improvements, since the 0-point is not necessarily an exact classification boundary, as shown in Fig. 9. Also, more complex prompts can be used. However, due to space constraints, we have not attempted this incremental approach, remaining it for an empirically possible practice.

We leave the efficiency benefit mentioned in §4.3 un-explained due to the consideration of length and complexity, also, observation in §5.2 needs more theoretical and experimental analysis. As we can see, some models (GPT2-XL) do not benefit from demonstrations even through the lens of hidden state aggregation or Hidden Calibration, which needs to be explained. An explanation of "why such aggregation occurs or not", and "how to enhance the intra-class aggregation by fine-tuning or prompt engineering" will be considerably beneficial for understanding ICL.

Acknowledgments

Not available during the anonymous review.

References

Momin Abbas, Yi Zhou, Parikshit Ram, Nathalie Baracaldo, Horst Samulowitz, Theodoros Salonidis, and Tianyi Chen. 2024. Enhancing in-context learning via linear probe calibration. *arXiv preprint arXiv:2401.12406*.

AI@Meta. 2024. Llama 3 model card.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A 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.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2022. Prototypical calibration for few-shot learning of language models. In *The Eleventh*

International Conference on Learning Representations.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. 2023. Generative calibration for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2312–2333.

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*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024. 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.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022c. 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.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

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.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2022. k nn prompting: Beyond-context learning with calibration-free nearest neighbor inference. In *The Eleventh International Conference on Learning Representations*.

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yufeng Zhao, Yoshihiro Sakai, and Naoya Inoue. 2024. Noisyicl: A little noise in model parameters calibrates in-context learning. *arXiv preprint arXiv:2402.05515*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2023. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*.

A Related Works

Given the topic of enhancing in-context learning, there are 3 classes of methods focused on such a target.

Model parameter update-based method: Although it is pointed out that the ICL objective is implicitly included in pre-training data (Han et al., 2023), explicitly fixing the gap between the ICL objective and causal language modeling objective can still be beneficial. Such methods are usually based on supervised fine-tuning (Min et al., 2022b; Gu et al., 2023; Wei et al., 2021, 2023; Iyer et al., 2022; Wang et al., 2022), and also self-supervised training (Chen et al., 2022) and non-gradient method (Zhao et al., 2024). Such methods usually require huge amounts of computation and data overhead to update billions of LM parameters.

In contrast, lightweight solutions focus on **Classification criteria-based method (calibration)**: Such methods focus on calculating output label logits, keeping the main feed-forward calculation processes and their parameters un-modified. The original motivation for these works is to eliminate prior bias and unfaithful confidence in ICL, by calibrating the output label probabilities (Holtzman et al., 2021; Shi et al., 2022; Fei et al., 2023; Zhao et al., 2021; Han et al., 2022; Zhou et al., 2023; Jiang et al., 2023). While, as described in the main text, some practices without the usage of label probabilities have also been proposed (Xu et al., 2022; Abbas et al., 2024; Min et al., 2022a).

Also, a careful **design of input prompts** can help improve the ICL performance. (1) Demonstration selection. Gonen et al. (2023) finds that selecting the demonstrations with lower perplexity improves the ICL performance, while similarly Kim

et al. (2022) generate the demonstrations from pre-trained LMs, etc. (2) Demonstration ordering. It is found that the ordering of demonstrations can significantly influence the performance, as also shown in our experiments in Fig. 5 (Lu et al., 2022; Liu et al., 2024; Xu et al., 2024). Specifically, Lu et al. (2022) detect the optimal demonstration ordering by some synthetic detecting sequences, while Liu et al. (2024) orders the demonstrations from easy to hard, following a curriculum learning form.

B Experimental Details

B.1 Datasets

In this paper, 10 datasets are used as shown in Table 1. Some datasets do not provide valid splitting, so we randomly split all of them into calibration sets and test sets: For each dataset, we first shuffle it with random seed 42. Then, we choose the 512 data at the tail as the testing data, and the 512 data at the head (all the datasets have more than 1024 examples) as the calibration data. Each data point in a test set is used once for each experiment trial to build a prompt example and test for performance.

`AGNews` and `GLUE-RTE` have over-length examples. So, in the main experiments, we filter out those examples: for LLaMa 2-6.9B, when $k = 8$, we filter out all the examples with a string length greater than 512 in `AGNews` and 128 in `GLUE-RTE`. Also, for LLaMa 3-8B, when $k = 8$, we filter out all the examples with a string length greater than 128 in GLUE-RTE and omit the experiments on `AGNews`. In the experiments in §5.2, for all the models, we filter out all the examples with a string length greater than 256 for all the k.

B.2 Baselines

6 baselines (1 vanilla and 5 improved) are used in this paper. Here we introduce the 5 calibration baseline.

Contextual Calibration (Con.C). Proposed by Zhao et al. (2021), Con.C uses empty queries with normal demonstrations as calibration samples. We input m samples with empty queries into the model and get the averaged normalized label probabilities $\bar{p}'$ among m samples. We take the reciprocal of the probabilities as calibration term $A = \bar{p}'^{-1}$, while the $B = \mathbf{0}$.

Batch Calibration (Bat.C). Proposed by Zhou et al. (2023), Bat.C is an inference-time calibration, using the negative averaged normalized label probabilities $-\bar{p}$ of m samples in inference time as the calibration term $B = -\bar{p}$, while the $A = \mathbf{1}$, where $\mathbf{1}$ is the all-one vector.

Domain Calibration (Dom.C). Proposed by Fei et al. (2023), Dom.C acts similarly to the Con.C. The difference is that it uses a random sequence sampled on the random tokens from the calibration dataset as queries instead of empty ones. We fix the sampled length to 32.

KNN Prompt (KNN). Proposed by Xu et al. (2022), KNN uses the whole output vocabulary probability distribution as the classification feature, instead of the label tokens. First, features of calibration examples are calculated as k-NN anchors. Then, during the inference, a k-NN classifier is used to classify the feature from the test samples. We use m examples per class to calculate the anchors for k-NN, and the nearest neighbor number is set to 3.

Central Calibration (Cent.C). This is the control method proposed by us. The calculation process is completely consistent with the Hidden Calibration, except that the hidden state is not used, and the whole output vocabulary probability distribution consistent with KNN is used. This method compares with Hidden Calibration to prove that the output probability distribution is not a good classification feature for ICL. ‘

Notice that: these label-probability-based methods (Con.C, Bat.C, Dom.C) use A or B *along*, which may be another major drawback of these calibration methods: According to Fig. 2, if a calibration rotates the mapping direction suitably, and transfer the 0-point properly, a decision boundary close to the Hidden Calibration can be found. This also leads to a new research direction for calibration: the simultaneous usage of translation and rotation methods.

B.3 Prompts

In this paper, we use a minimum prompt template shown in Table 2.

To facilitate the replication of label probability-based methods, we limit the label space to one token by synonymous conversion. Note that Hidden Calibration does not need to meet such a one-token requirement. That is, Hidden Calibration can be applied to classification datasets of any length on the label.

Table 1: Datasets and Abbreviations used in this paper.

Dataset	Abbr.
`AGNews` (Zhang et al., 2015)	AGNews
`SemEval 2014-Task 4 Restaurant` (Pontiki et al., 2014)	SemE.R
`SemEval 2014-Task 4 Laptops` (Pontiki et al., 2014)	SemE.L
`Poem Sentiment` (Sheng and Uthus, 2020)	PoemS
`GLUE-RTE` (Wang et al., 2019)	RTE
`tweet_eval_emotion` (Mohammad et al., 2018)	TEE
`tweet_eval_hate` (Basile et al., 2019)	TEH
`tweet_eval_sentiment` (Rosenthal et al., 2017)	TES
`financial_phrasebank (all agree)` (Malo et al., 2014)	FP
`rotten_tomatoes` (Pang and Lee, 2005)	Rott.T

Table 2: Prompt templates used in this paper.

Dataset	Prompt Template	Verbalizer
AGNews	Input: <x>, Label: <y>	world, sport, business, science
SemE.R	Input: <x>, Aspect: <a>, Label: <y>	positive, neutral, negative
SemE.L	Input: <x>, Aspect: <a>, Label: <y>	positive, neutral, negative
PoemS	Input: <x>, Label: <y>	positive, neutral, negative, mix
RTE	Input: <x>, Text 2: <a>, Label: <y>	include, neutral
TEE	Input: <x>, Label: <y>	anger, joy, positive, sad
TEH	Input: <x>, Label: <y>	normal, hate
TES	Input: <x>, Label: <y>	positive, neutral, negative
FP	Input: <x>, Label: <y>	positive, neutral, negative
Rott.T	Input: <x>, Label: <y>	positive, negative

Especially, in §4.3, we use 6 more prompt templates to test the stability of each ICL method against the prompt templates. We list these extra templates

B.4 Details of Visualization in §3.1

Principle Component Analysis (PCA). Given a hidden state set $\mathcal{H} = \left\{h^{(i)}\right\}_{i=1}^{n}$, we span all the hidden state vector into a matrix $H \in \mathbb{R}^{n\times d}$. The covariance matrix is $\operatorname{cov}(H) = \frac{1}{n}\left(H-\bar{H}\right)^T\left(H-\bar{H}\right)$, where the $\bar{H}$ is the matrix spanned by the element-wise average vectors $\bar{h}$ of hidden state set $\mathcal{H}$. We conduct Eigenvalue Decomposition on $\operatorname{cov}(H)$ and adjust the dimensions to arrange the eigenvalues Λ in a descending order along the row:

$$\operatorname{cov}(H) = P\Lambda P^T, \tag{1}$$

where the $P \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Taking the top-$\tilde{d}$ lines of P and span them into $\tilde{P} \in \mathbb{R}^{d\times\tilde{d}}$, we get the principle component mapping:

$$\mathrm{PCA}_{\mathcal{H}}(h) = \left(h-\bar{h}\right)\tilde{P} = h\tilde{P} - \bar{h}\tilde{P}. \tag{2}$$

Notice that $\tilde{P}\tilde{P}^T = I$, where I is the identity matrix.

Dot-product after PCA. Suppose we have dot-product with vector[4] h and E in the original space $\mathbb{R}^d$, producing the dot-product similarity classification criterion α:

$$\alpha = h\left(E^T - \mathbf{0}^T\right). \tag{3}$$

When we conduct a same PCA on both h and E^T to get dot-product similarity in a dimensionality-reduced space similar to Fig. 2:

$$\tilde{\alpha} = \mathrm{PCA}_{\mathcal{H}}(h)\underbrace{\left(\mathrm{PCA}_{\mathcal{H}}(E)^T - \mathrm{PCA}_{\mathcal{H}}(\mathbf{0})^T\right)}_{\text{Mapping direction selected after PCA}} \tag{4}$$

$$= \left(h\tilde{P} - \bar{h}\tilde{P}\right)\left(E\tilde{P}\right)^T \tag{5}$$

$$= h\tilde{P}\tilde{P}^TE^T - \bar{h}\tilde{P}\tilde{P}^TE^T \tag{6}$$

$$= \alpha - \bar{h}E^T. \tag{7}$$

Notice that we use the mapping direction $\left(\mathrm{PCA}_{\mathcal{H}}(E)^T - \mathrm{PCA}_{\mathcal{H}}(\mathbf{0})^T\right)$ after the PCA, instead of $\left(\mathrm{PCA}_{\mathcal{H}}(E)^T - \mathbf{0}^T\right)$, and this is the reason why the oblique axis in Fig. 2 does not necessarily pass through the coordinate origin. In such a scenario, the dot productions after PCA only differ by a fixed constant bias $-\bar{h}E^T$ from the ones before PCA. This is the reason why the normal line of oblique axis on the 0-point doesn't pass the coordinate origin of the 2D-plane in Fig. 2.

[4]Due to excessive superscripts, in this section, we omit the superscripts U in the notation of un-embedding E_l^U.

Table 3: The 7 prompt templates used in the experiment (on `Rotten_Tomatoes`) of Fig. 5 (left).

#	Prompt Template
Original	Input: <x>, Label: <y>
1	Sentence: <x>, Label: <y>
2	sentence: <x>, Label: <y>
3	sentence:\n <x>, Label: <y>
4	Input: <x>, Sentiment: <y>
5	Input: <x>, sentiment: <y>
6	x: <x>, y: <y>

Table 4: The **additional** (compare to vanilla ICL) time and space on calibration and inference cost of various methods. Hidden Calibration has a similar cost upper bound to other calibrations. $|\mathbb{V}|$ is the vocabulary size.

Method	Calibration Cost		Inference Cost
	Add. Space	Add. Time	Add. Time
None	0	0	0
Con.C	$O(\lvert\mathcal{Y}\rvert)$	$O(m)$	$O(\lvert\mathcal{Y}\rvert)$
Bat.C	0	0	$O(m\lvert\mathcal{Y}\rvert)$
Dom.C	$O(\lvert\mathcal{Y}\rvert)$	$O(m)$	$O(\lvert\mathcal{Y}\rvert)$
KNN	$O(m\lvert\mathbb{V}\rvert)$	$O(m)$	$O(m\lvert\mathbb{V}\rvert)$
Cent.C	$O(\lvert\mathcal{Y}\rvert\lvert\mathbb{V}\rvert)$	$O(m)$	$O(\lvert\mathcal{Y}\rvert\lvert\mathbb{V}\rvert)$
Hidd.C	$O(\lvert\mathcal{Y}\rvert d)$	$O(m)$	$O(\lvert\mathcal{Y}\rvert d)$

Decision Boundary after PCA. Notice that the decision boundary of two classes l_1 and l_2 in an non-rotated ICL scenario is:

$$\mathcal{B} = \left\{h | hE_{l_1}^T - hE_{l_2}^T = C\right\}. \tag{8}$$

Where the C is the calibration term without rotation. Notice that it is a hyperplane in $\mathbb{R}^d$ with normal vector $(E_{l_1} - E_{l_2})^T$. Also, the normal plane which pass the 0-point of direction $(E_{l_1} - E_{l_2})^T$ in $\mathbb{R}^{\tilde{d}}$ after PCA is:

$$\begin{aligned}\tilde{\mathcal{B}} = \{&\mathrm{PCA}_{\mathcal{H}}(h) | \mathrm{PCA}_{\mathcal{H}}(h) \\ &(\mathrm{PCA}_{\mathcal{H}}(E_{l_1} - E_{l_2}) - \mathrm{PCA}_{\mathcal{H}}(\mathbf{0}))^T = 0\}.\end{aligned} \tag{9}$$

By the aforementioned transformation, we have:

$$\tilde{\mathcal{B}} = \left\{\mathrm{PCA}_{\mathcal{H}}(h) | hE_{l_1}^T - hE_{l_2}^T = \bar{h}\left(E_{l_1}^T - E_{l_2}^T\right)\right\}. \tag{10}$$

That is, the dimensionality-reduced decision boundary $\tilde{\mathcal{B}}$ is perpendicular to the mapped direction $(\mathrm{PCA}_{\mathcal{H}}(E_{l_1} - E_{l_2}) - \mathrm{PCA}_{\mathcal{H}}(\mathbf{0}))$, and biased only by a constant $\left(\bar{h}\left(E_{l_1}^T - E_{l_2}^T\right) - C\right)$ on the classification criteria comparing to the original space. Specifically, in the two-dimensional case, it is a straight line that may not necessarily pass through the coordinate origin, as shown in Fig. 2.

B.5 Details of Experiment in §5.1

B.5.1 Calculation Details of Averaged Overlap

First, we divide the $|\mathcal{Y}|$-way classification task into $\mathbb{C}(|\mathcal{Y}|, 2)$ 2-way classification task[5], to allow us to use a scalar to characterize the classification criteria for each 2-combination (similar to what we do to the "positive" and "negative" examples in Fig. 2). Then, for each chosen 2-combination, w.l.o.g, given labels denoted as l_1 and l_2, we build prompt-label sets[6] as:

$$\mathcal{S}_{l_j} = \left\{T\left(\mathcal{D}^{de,(i)}, x^{(c_i)}\right) \middle| y^{(c_i)} = l_j\right\}_{l_j \in \{l_1, l_2\}}^{n_{l_j}}, \tag{11}$$

where c_i is the sampled query index. That is, we sample queries annotated with these two labels and build prompt sets, then collect the prompts with the same query label l_j into $\mathcal{S}_{l_j}$, with a size n_{l_j}.

Then, for each prompt $s^{(i)} = T\left(\mathcal{D}^{de,(i)}, x^{(i)}\right) \in \mathcal{S}_{l_j}$, we run decoders (vanilla, Con.C, Dom.C and Hidden Calibration) with probability normlization $f_{l_1}(\cdot)$ and $f_{l_2}(\cdot)$ to get the classification probabilities of assigning label l_1 and l_2 as $\alpha_1^{(i)} = f_{l_1}\left(s^{(i)}\right)$ and $\alpha_2^{(i)} = f_{l_2}\left(s^{(i)}\right)$. We calculate the difference between $\alpha_1^{(i)}$ and $\alpha_2^{(i)}$ and collect them into a set:

$$\mathcal{A}_{l_j} = \left\{\alpha_1^{(i)} - \alpha_2^{(i)} \middle| s^{(i)} \in \mathcal{S}_{l_j}\right\}_{i=1}^{n_{l_j}}. \tag{12}$$

Now, for the 2-combination of labels (l_1, l_2), we get $\mathcal{A}_{l_1}$ and $\mathcal{A}_{l_2}$, whose elements are the probabilities difference between assigning l_1 and assigning l_2 to example $s^{(i)}$. The difference between $\mathcal{A}_{l_1}$ and $\mathcal{A}_{l_2}$ is: the elements in $\mathcal{A}_{l_1}$ are from $s^{(i)}$s with queries labeled by ground-truth l_1, and vice versa. We obtain continuous probability density functions of $\mathcal{A}_{l_1}$ and $\mathcal{A}_{l_2}$ as $p_{l_1}(\cdot)$ and $p_{l_2}(\cdot)$ by kernel density estimation, as the curves in Fig. 7.

Then, we calculate the overlap area of these curves:

$$S_{l_1,l_2} = \int_{-1}^{1} \min\left[p_{l_1}(x), p_{l_2}(x)\right] \mathrm{d}x. \tag{13}$$

For each combination[7] in the $\mathbb{C}(|\mathcal{Y}|, 2)$ 2-combinations, we repeat to calculate the $S_{\cdot,\cdot}$, and average them as the **Averaged Overlap** $\bar{S}$.

$$\bar{S} = \frac{1}{\mathbb{C}(|\mathcal{Y}|, 2)} \sum_{i=1}^{|\mathcal{Y}|} \sum_{j=i+1}^{|\mathcal{Y}|} S_{l_i,l_j}. \tag{14}$$

B.5.2 Experimental Details

We conduct experiments resulting Fig. 8 on 3 models with `SemEval 2014-Task 4 Restaurant`, `SemEval 2014-Task 4 Laptops`, AGNews, Poem Sentiment, and fiancial_phrasebank, given the demonstration number $k = 4$ and calibration example numbers $m = 16$. We use the whole 512 examples on the test split for each dataset and repeat 5 times.

B.5.3 Proof: *the Overlap Area is Double to the Error's Lower Bound*

Suppose a label combination l_1 and l_2, w.l.o.g., we have a ground truth probability density function $p_{l_1}(x)$ and $p_{l_2}(x)$ on a criterion $x \in \mathbb{X}$, same as the curves in Fig. 7. Given a specific value of criterion x, the upper-bound classification performance is determined by majority vote, which is the most accurate method on such a point, resulting in a density of error classification:

$$e(x)_{l_1,l_2} \geqslant \min\left[p_{l_1}(x), p_{l_2}(x)\right]. \tag{15}$$

[5] The $\mathbb{C}(m, n)$ is the n-combination number from m elements.

[6] Notice that the T is the prompting function.

[7] Notice that on $S_{\cdot,\cdot}$, the labels are rotational symmetry.

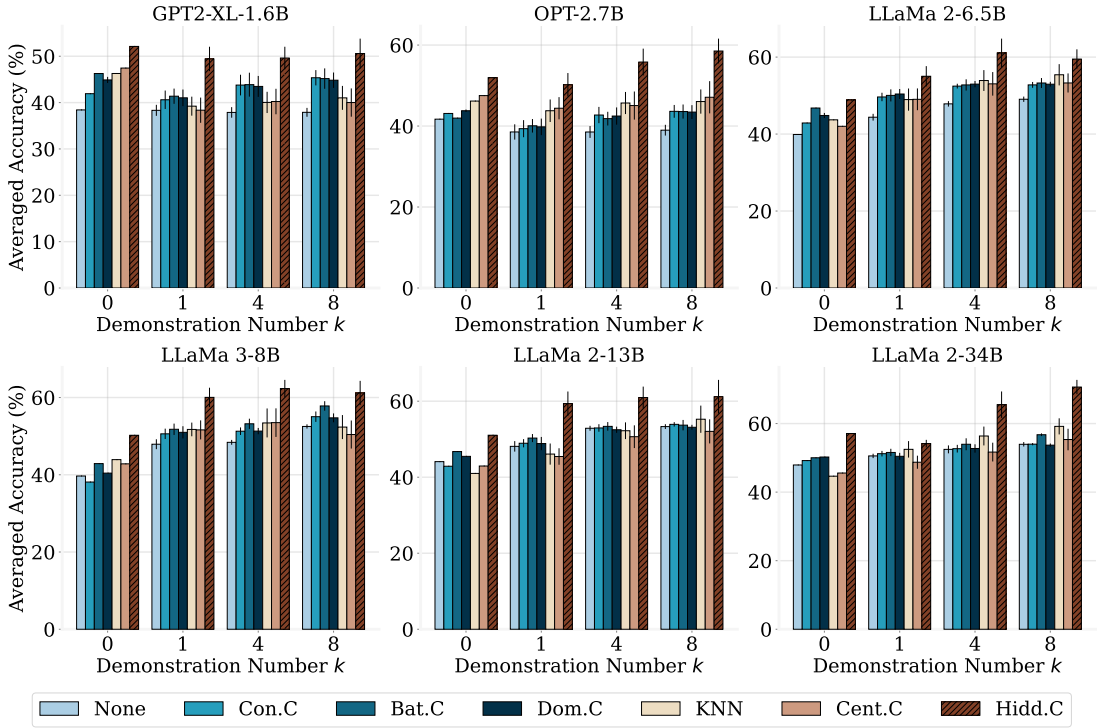

Figure 13: The classification performance (Accuracy(%)) of 3 models averaged on 10 datasets.

Table 5: Transferability of centroid among various datasets with the same label space. Big numbers are the averaged improvement (MF1) compared to vanilla ICL, small numbers are standard error. Statistically significant results ($p < 0.1$) are in **bold**.

Test \ Cali.	SemE.R	SemE.L	Fina.P	TES
SemE.R	**(+38.75)** ±2.28	**+29.24** ±3.19	+6.32 ±10.55	+7.54 ±8.96
SemE.L	**+20.78** ±7.37	**(+37.33)** ±3.47	-0.40 ±7.37	+8.94 ±8.93
Fina.P	**+7.42** ±4.98	+9.05 ±11.14	**(+37.29)** ±2.30	-4.35 ±6.34
TES	+6.95 ±7.00	**+9.73** ±5.68	-0.51 ±3.83	**(+11.83)** ±3.59

Table 6: Transferability of centroid among various k on the same dataset. $k_1 \rightarrow k_2$ is to use centroids estimated by k_1 demonstrations for inference on test examples with k_2 demonstrations. Other annotations are the same as Table. 5

	0→1	4→1	(1→1)	1→4	(4→4)
SemE.R	**+9.46** ±1.95	**+22.50** ±14.55	**(+26.14)** ±5.16	**+17.95** ±7.51	**(+38.75)** ±2.28
SemE.L	**+26.80** ±3.20	**+17.18** ±5.61	**(+26.65)** ±2.72	+10.79 ±14.86	**(+37.33)** ±3.47
AGNews	**+42.38** ±2.42	**+40.20** ±1.24	**(+41.02)** ±2.49	**+43.12** ±2.02	**(+46.66)** ±3.77
PoemS	+0.16 ±1.87	+2.12 ±6.18	**(+21.49)** ±2.54	**+8.79** ±1.84	**(+12.96)** ±1.52
Fina.P	-0.13 ±1.88	**+21.40** ±2.90	**(+16.70)** ±3.80	**+10.00** ±13.68	**(+37.30)** ±2.30

So, the integral error rate:

$$\mathcal{E}_{l_1,l_2} \geqslant \frac{\int_{x\in\mathbb{X}} \min\left[p_{l_1}(x), p_{l_2}(x)\right] \mathrm{d}x}{\int_{x\in\mathbb{X}} p_{l_1}(x)dx + \int_{x\in\mathbb{X}} p_{l_2}(x)\mathrm{d}x} \quad (16)$$

$$= \frac{1}{2}\int_{x\in\mathbb{X}} \min\left[p_{l_1}(x), p_{l_2}(x)\right] \mathrm{d}x \quad (17)$$

$$= \frac{1}{2} S_{l_1,l_2}. \quad (18)$$

B.6 Details of Experiment in §5.2

B.6.1 Calculation of the Distance and Standard Error

Averaged Centroid Distance. Given a $|\mathcal{Y}|$-way classification task, for each label l we build its corresponding prompt set $\mathcal{S}_l = \left\{s^{(c_i)} | y^{(c_i)} = l\right\}_{i=1}^{n_l}$, where $s^{(c_i)}$ is the prompt with query labeled by l, and c_i is the sampled query index. We encode it into a hidden state set $\mathcal{H}_l = \left\{h^{(i)}\right\}_{i=1}^{n_l}$, and calculate its centroid $\bar{h}_l$, as what we do in Hidden Calibration:

$$\bar{h}_l = \frac{1}{n_l} \sum_{h^{(i)} \in \mathcal{H}_l} h^{(i)}. \quad (19)$$

For every 2-combination of labels l and l', we calculate the distance of their centroid, and the average among all the 2-combination is used as the Aver-

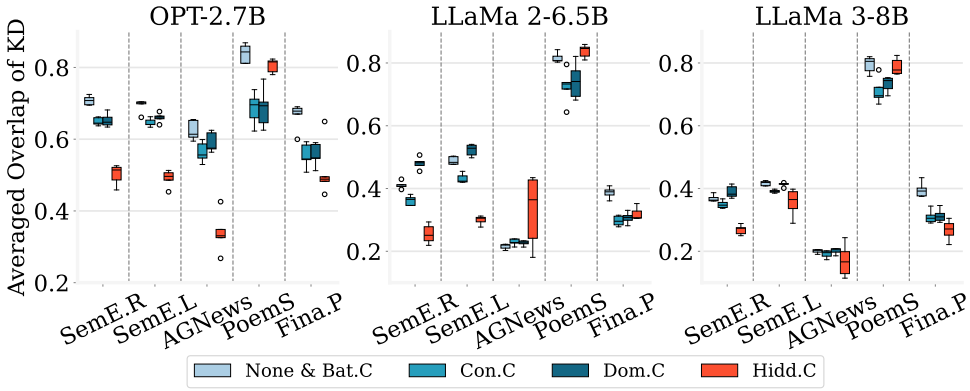

Figure 14: The augmented results on 2 models of Fig. 8.

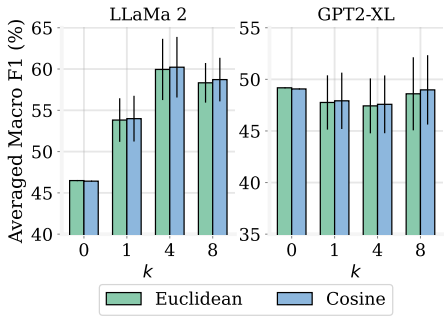

Figure 15: The classification performance (Macro F1(%)) of Hidden Calibration with difference similarity measure.

aged Centroid Distance:

$$\mathrm{ACD} = \frac{1}{\mathbb{C}(|\mathcal{Y}|, 2)} \sum_{i=1}^{|\mathcal{Y}|} \sum_{j=i+1}^{|\mathcal{Y}|} \left\| \bar{h}_i - \bar{h}_j \right\|_2 . \quad (20)$$

Averaged Intra-class Standard Error. Given the hidden state set $\mathcal{H}_l = \left\{h^{(i)}\right\}_{i=1}^{n_l}$ w.r.t. the label l, we span all the hidden state vectors into a matrix $H_l \in \mathbb{R}^{n_l \times d}$. The covariance matrix is $\left(H_l - \bar{H}_l\right)^T \left(H_l - \bar{H}_l\right)$, where the $\bar{H}_l$ is the matrix spanned by the element-wise average vectors of hidden state set $\mathcal{H}_l$. Notice that the ACD is a first-order moment, for a proper comparison, we use the average on the diagonal elements of the element-wise square root of the covariance matrix as the intra-class standard error metric for label l. We average all the standard errors from all the classes as the Averaged Intra-class Standard Error:

$$\mathrm{AIS} = \frac{1}{|\mathcal{Y}| d} \sum_{i=1}^{|\mathcal{Y}|} \mathrm{tr} \left[\sqrt{\left(H_i - \bar{H}_i\right)^T \left(H_i - \bar{H}_i\right)} \right] . \quad (21)$$

B.6.2 Experimental Details

We conduct experiments resulting Fig. 12 on 4 models with `SemEval 2014-Task 4 Restaurant`, `SemEval 2014-Task 4 Laptops`, AGNews, Poem Sentiment, and fiancial_phrasebank, given the calibration example numbers $m = 16$. We use the whole 512 examples on the test split for each dataset and repeat 5 times.

B.7 Experimental Details for Fig. 6

We conduct experiments resulting Fig. 6 on OPT-2.7B with 4 datasets: `SemEval 2014-Task 4 Restaurant`, `SemEval 2014-Task 4 Laptops`, AGNews, and Poem Sentiment, given the demonstration numbers $k = 4$ and repeat 5 times.

C Detailed Results

C.1 Details of Main Results

Numerical details of Fig. 4 are shown in Table 8, 9, 10, 11, 12 and 13. Accuracy results are shown in Fig. 13.

C.2 Details of Averaged Overlaps Results

The augmented results on the other 3 models (we skip this experiment on the quantitated model) of Fig. 8 are shown in Fig. 14.

D Additional Discussion

D.1 The Similarity Measures Used in Hidden Calibration

In §3.2, we use the Euclidean distance as the similarity measure. But this is not the only option. Intuitively, we can choose other similarity measures as alternatives. Since we get inspired by observation with dot-production similarity, we have an obligation to check the performance on such a measure instead of the Euclidean distance. This section uses

Table 7: Performance of all 6 models on TREC for a more-way classification, and on Hate_Speech18 for a biased dataset. $k = 4$, top-2 results are in **bold**.

	GPT2-XL-1.6B	OPT-2.7B	LLaMa 2-6.5B	LLaMa 3-8B	LLaMa 2-13B	LLaMa 2-34B
			TREC			
None	13.02	15.85	23.12	16.79	23.22	21.13
Con.C	14.10	8.12	23.72	18.34	23.56	22.35
Bat.C	14.44	17.31	23.26	20.28	22.88	22.15
Dom.C	14.10	8.58	23.14	19.10	23.42	23.23
KNN	**27.53**	33.80	49.25	43.15	54.76	**53.66**
Cent.C	23.46	**33.80**	**49.18**	**46.12**	**56.79**	53.03
Hidd.C	**55.91**	**61.14**	**64.90**	**71.59**	**75.64**	**68.39**
			Hate_Speech18			
None	**23.96**	23.95	**23.96**	**23.96**	**23.96**	23.92
Con.C	21.57	25.47	14.68	**23.74**	21.49	21.25
Bat.C	20.59	21.25	20.54	18.25	20.66	23.68
Dom.C	21.23	23.68	20.73	23.67	**24.36**	23.28
KNN	17.30	16.74	16.50	16.26	16.23	**27.05**
Cent.C	16.32	**25.47**	15.42	5.46	9.15	22.23
Hidd.C	**24.47**	**27.03**	**23.44**	20.67	20.06	**29.94**

cosine similarity as an example to illustrate that there is no significant performance difference between these measures. We use cosine similarity to repeat the results in §4.2 on LLaMa 2-6.9B and GPT2-XL.

The results are shown in Fig. 15, where the performance based on these two measures is close, without statistical difference. This indicates that the hidden space has good properties of both metric and vector space, and Hidden Calibration acts equally on these measures.

D.2 Time and Space Cost of Hidden Calibration and baselines

We analyze the upper bound of the additional space-time cost of the baseline method and Hidden Calibration, as shown in Table 4. Here, we are most concerned about the inference time cost, and Hidden Calibration is the fastest among all the non-label-based methods.

Since the product $|\mathcal{Y}|d$ is usually not very large, Hidden Calibration does not add considerable inference overhead. In contrast, KNN may be incredibly slow as the calibration dataset scales.

D.3 Transferability of the Centorid

We have proven that it is not advisable to use the *common* token probability criteria, while, since the centroid criteria are proven to be better than token probability, we are curious: can the centroid calculated in one task be transferred to other tasks with the same label space? Among the datasets sharing the same label space "positive", "neutral", and "negative", we calculate centroids by one dataset and evaluate Hidden Calibration with it on another dataset, on OPT-2.7B, with $k = 4$, $m = 16|\mathcal{Y}|$. The results are shown in Table 5, where only limited transferability is demonstrated in different domains of the same task (SemE.R and SemE.L), whose behavior is similar to *task vector* (Ilharco et al., 2022; Hendel et al., 2023), while other combination of datasets can not demonstrate considerable transferability. This further exacerbates our doubts about the token-based method: We find that the hidden state distributions have significant differences among various datasets, even if they share a common label space, then utilizing fixed token unembedding vectors to decode these classification criteria is highly unreliable.

Moreover, we repeat this experiment on various k, instead of various datasets, as shown in Table 6. The transferabilities among k are better than on datasets, but still worse than the un-transferred scenario. Notice that $4 \to 1$ results are much better than $0 \to 1$, which support our results in §5.2: hidden states with higher k are further converged.

D.4 A Demonstration towards ICL Principles

Our findings may lead to an explanation of the principle of ICL and traditional calibrations. LMs generate distributed representations into separate clusters in the last hidden state. At this point, by dot-product, any non-collinear *arbitrary or plausible* mapping directions should be able to capture and classify these clusters to some extent. Note: The absolute distance in such a direction is not

faithful (since the centroids of these hidden states and the coordinate origins in these mapping directions are not necessarily aligned), which leads to the generation of so-called bias, and calibrating these biases can improve the performance to a certain extent. However, in such a paradigm, high-dimensional features are discarded, resulting in overlapping originally linearly separable features in high-dimensional space, leading to a loss of classification accuracy, even if the calibration aligns the coordinate origin.

D.5 Applicability on More-way Classification and Biased Dataset

To further verify the applicability of Hidden Calibration on harder tasks, we test the performance of Hidden Calibration and baseline methods on TREC (Li and Roth, 2002; Hovy et al., 2001) for more-way ($|\mathcal{Y}| = 6$) classification and Hate_Speech18 (de Gibert et al., 2018) for a biased dataset (label frequency distribution: $[0.87, 0.11, 0.007, 0.015]$), shown in Table 7. For results on TREC, Hidden Calibration produces a significant improvement compared to all the baselines on all the models. While, Hidden Calibration outperforms on Hate_Speech18 in most cases, and sometimes Hidden Calibration is weaker than baseline methods, but consistently produces competitive results. However, we believe that this slightly weaker result cannot be fully attributed to biased datasets, given that there are biased datasets among the 10 standard datasets where Hidden Calibration perform SotA results (e.g. `SemEval 2014-Task 4 Restaurant`, refer §C.1).

E Statements

E.1 Author Contributions Statement

Not available during the anonymous review.

E.2 License for Artifacts

Models. GPT2-XL and OPT-2.7B is under the MIT license, LLaMa family is under its specific license.

Datasets. We list the open-source license for the datasets used in this paper as follows:

- `CC-by-4.0`: `Poem Sentiment`, `SemEval 2014-Task 4 Restaurant`, `SemEval 2014-Task 4 Laptops`, `tweet_eval_emotion`, `tweet_eval_hate`, `tweet_eval_hate`
- `CC-by-SA-3.0`: `financial_phrasebank`, `GLUE-RTE`
- Unknown: `AGNews`, `rotten_tomatoes`

Consistency of Usage. Models and data are used with their original usage.

E.3 AI Agent Usage

AI Agents are only used for writing improving and grammar checking in this paper.

Table 8: Classification performance (Macro F1(%)) on GPT2-XL. mean_{std}, top-2 results are in **bold**.

GPT-2 XL		AGNews	SemE.R	SemE.L	PoemS	RTE	TEE	TEH	TES	Fina.P	Rott.T	Average
k=0	None	$16.53_{0.00}$	$9.87_{0.00}$	$12.31_{0.00}$	$8.75_{0.00}$	$48.31_{0.00}$	$19.40_{0.00}$	$37.56_{0.00}$	$21.14_{0.00}$	$25.36_{0.00}$	$34.16_{0.00}$	23.34
	Con.C	$30.04_{0.00}$	$9.87_{0.00}$	$12.31_{0.00}$	$8.11_{0.00}$	$47.46_{0.00}$	$9.68_{0.00}$	$42.97_{0.00}$	$27.01_{0.00}$	$33.14_{0.00}$	$60.75_{0.00}$	28.13
	Bat.C	$42.56_{0.00}$	$37.29_{0.00}$	$\mathbf{46.21}_{0.00}$	$18.92_{0.00}$	$\mathbf{51.29}_{0.00}$	$26.39_{0.00}$	$\mathbf{44.45}_{0.00}$	$\mathbf{42.52}_{0.00}$	$32.00_{0.00}$	$\mathbf{66.96}_{0.00}$	40.86
	Dom.C	$38.00_{0.60}$	$24.76_{0.04}$	$26.58_{1.06}$	$20.52_{0.99}$	$38.00_{0.60}$	$14.34_{0.00}$	$37.56_{0.00}$	$\mathbf{42.70}_{1.00}$	$30.14_{1.06}$	$\mathbf{67.42}_{2.19}$	34.33
	KNN	$50.63_{0.00}$	$\mathbf{39.62}_{0.00}$	$42.31_{0.00}$	$26.69_{0.00}$	$\mathbf{49.57}_{0.00}$	$30.79_{0.00}$	$\mathbf{44.68}_{0.00}$	$34.12_{0.00}$	$39.30_{0.00}$	$62.37_{0.00}$	42.01
	Cent.C	$\mathbf{52.54}_{0.00}$	$39.53_{0.00}$	$45.37_{0.00}$	$\mathbf{27.23}_{0.00}$	$44.92_{0.00}$	$\mathbf{32.71}_{0.00}$	$41.93_{0.00}$	$32.15_{0.00}$	$\mathbf{43.15}_{0.00}$	$63.32_{0.00}$	**42.29**
	Hidd.C	$\mathbf{82.02}_{0.00}$	$\mathbf{44.73}_{0.00}$	$\mathbf{54.45}_{0.00}$	$\mathbf{32.81}_{0.00}$	$47.36_{0.00}$	$\mathbf{43.45}_{0.00}$	$42.95_{0.00}$	$36.78_{0.00}$	$\mathbf{47.38}_{0.00}$	$59.83_{0.00}$	**49.18**
k=1	None	$20.95_{1.29}$	$36.72_{1.19}$	$31.60_{1.33}$	$21.21_{1.72}$	$49.47_{2.22}$	$22.81_{0.94}$	$37.56_{0.00}$	$31.74_{1.43}$	$30.15_{1.97}$	$36.20_{0.56}$	31.84
	Con.C	$24.15_{1.13}$	$41.88_{1.73}$	$38.92_{3.32}$	$24.82_{2.55}$	$47.64_{3.38}$	$20.70_{1.20}$	$37.56_{0.00}$	$33.06_{1.29}$	$34.93_{1.31}$	$60.21_{3.98}$	36.39
	Bat.C	$30.02_{1.49}$	$\mathbf{45.04}_{0.77}$	$\mathbf{41.10}_{4.41}$	$25.20_{0.48}$	$\mathbf{49.58}_{1.88}$	$25.85_{2.51}$	$\mathbf{48.02}_{0.94}$	$\mathbf{34.92}_{0.91}$	$\mathbf{35.02}_{1.14}$	$\mathbf{64.72}_{1.92}$	**39.95**
	Dom.C	$22.17_{1.15}$	$44.86_{1.16}$	$41.02_{4.91}$	$\mathbf{25.61}_{1.27}$	$46.81_{2.09}$	$18.92_{1.58}$	$37.56_{0.00}$	$\mathbf{33.96}_{0.90}$	$34.72_{1.56}$	$62.27_{3.33}$	36.79
	KNN	$\mathbf{32.14}_{0.85}$	$36.88_{2.76}$	$37.29_{3.20}$	$21.44_{1.57}$	$48.18_{2.09}$	$\mathbf{27.44}_{1.75}$	$43.52_{3.73}$	$33.85_{1.08}$	$\mathbf{36.23}_{2.07}$	$52.41_{1.44}$	36.94
	Cent.C	$26.74_{2.18}$	$33.00_{1.48}$	$32.02_{3.63}$	$18.75_{2.78}$	$47.82_{3.27}$	$23.82_{3.95}$	$44.07_{1.52}$	$30.68_{2.12}$	$29.58_{1.98}$	$58.41_{4.30}$	34.49
	Hidd.C	$\mathbf{65.15}_{1.77}$	$\mathbf{49.16}_{3.43}$	$\mathbf{51.56}_{2.83}$	$\mathbf{32.83}_{2.32}$	$\mathbf{50.47}_{1.41}$	$\mathbf{36.17}_{2.74}$	$\mathbf{49.16}_{1.47}$	$33.55_{4.45}$	$44.02_{3.36}$	$\mathbf{65.55}_{2.47}$	**47.76**
k=4	None	$21.87_{4.32}$	$33.14_{1.46}$	$41.03_{2.14}$	$20.11_{1.47}$	$40.48_{1.15}$	$17.98_{0.33}$	$38.19_{1.41}$	$29.06_{1.70}$	$28.86_{2.07}$	$33.81_{0.72}$	30.45
	Con.C	$24.22_{10.00}$	$44.76_{0.98}$	$\mathbf{48.90}_{2.71}$	$21.95_{1.28}$	$36.33_{1.11}$	$24.40_{0.82}$	$37.51_{0.11}$	$\mathbf{37.30}_{2.49}$	$41.65_{2.96}$	$69.57_{1.11}$	38.66
	Bat.C	$26.97_{9.22}$	$44.48_{1.67}$	$46.94_{1.72}$	$21.93_{1.24}$	$47.17_{1.96}$	$\mathbf{28.86}_{1.88}$	$46.94_{2.76}$	$\mathbf{36.49}_{2.64}$	$\mathbf{44.82}_{2.54}$	$\mathbf{71.94}_{1.24}$	**41.65**
	Dom.C	$25.30_{10.23}$	$\mathbf{45.44}_{1.54}$	$47.01_{2.20}$	$\mathbf{23.05}_{1.05}$	$36.79_{1.68}$	$27.11_{1.72}$	$37.79_{0.51}$	$34.67_{0.81}$	$42.54_{2.53}$	$67.00_{3.89}$	38.67
	KNN	$\mathbf{33.93}_{2.04}$	$37.57_{3.54}$	$38.35_{2.35}$	$21.72_{3.05}$	$\mathbf{48.55}_{2.97}$	$26.20_{1.96}$	$\mathbf{48.71}_{2.74}$	$30.10_{2.12}$	$36.57_{2.56}$	$57.36_{1.72}$	37.91
	Cent.C	$32.98_{2.56}$	$37.24_{5.80}$	$32.71_{4.50}$	$18.47_{1.88}$	$45.78_{2.93}$	$24.21_{4.17}$	$\mathbf{48.83}_{3.57}$	$29.75_{2.50}$	$33.30_{7.56}$	$58.81_{2.86}$	36.21
	Hidd.C	$\mathbf{49.55}_{3.29}$	$\mathbf{50.81}_{2.16}$	$\mathbf{54.16}_{3.62}$	$\mathbf{24.96}_{2.56}$	$\mathbf{49.28}_{2.42}$	$\mathbf{39.13}_{2.80}$	$48.48_{1.62}$	$34.43_{2.46}$	$\mathbf{50.80}_{3.51}$	$\mathbf{72.70}_{2.19}$	**47.43**
k=8	None	$19.23_{0.78}$	$32.79_{2.13}$	$33.36_{1.02}$	$17.93_{1.08}$	$37.37_{2.07}$	$15.75_{0.75}$	$37.56_{0.00}$	$25.15_{0.99}$	$26.45_{1.13}$	$33.55_{0.45}$	27.91
	Con.C	$18.38_{0.28}$	$\mathbf{47.06}_{3.77}$	$\mathbf{52.40}_{0.96}$	$19.70_{2.05}$	$35.98_{0.00}$	$26.80_{2.00}$	$37.56_{0.00}$	$\mathbf{38.12}_{3.15}$	$41.05_{2.25}$	$\mathbf{75.05}_{2.12}$	39.21
	Bat.C	$22.32_{1.22}$	$45.96_{2.08}$	$48.25_{0.79}$	$\mathbf{21.20}_{1.55}$	$\mathbf{45.82}_{4.68}$	$\mathbf{29.23}_{1.22}$	$47.84_{1.36}$	$\mathbf{38.30}_{3.09}$	$\mathbf{48.78}_{2.06}$	$74.81_{1.58}$	**42.25**
	Dom.C	$21.85_{0.95}$	$45.78_{2.48}$	$49.91_{0.74}$	$20.80_{1.97}$	$35.98_{0.00}$	$28.50_{1.30}$	$37.56_{0.60}$	$36.20_{2.44}$	$47.99_{1.84}$	$69.87_{4.70}$	39.44
	KNN	$\mathbf{39.81}_{1.81}$	$38.83_{4.38}$	$34.38_{3.72}$	$20.92_{2.98}$	$43.03_{5.31}$	$26.49_{3.84}$	$\mathbf{49.00}_{1.23}$	$29.33_{3.74}$	$39.00_{1.91}$	$60.08_{3.64}$	38.09
	Cent.C	$41.52_{4.18}$	$37.73_{3.15}$	$32.21_{2.80}$	$14.66_{4.00}$	$45.11_{6.67}$	$21.83_{6.60}$	$48.97_{3.71}$	$25.81_{2.72}$	$31.32_{3.89}$	$55.23_{4.35}$	35.44
	Hidd.C	$\mathbf{57.10}_{5.54}$	$\mathbf{49.86}_{2.55}$	$\mathbf{58.26}_{3.15}$	$\mathbf{24.42}_{1.28}$	$\mathbf{48.48}_{4.48}$	$\mathbf{35.48}_{5.04}$	$\mathbf{51.50}_{2.18}$	$27.46_{6.30}$	$\mathbf{56.99}_{1.72}$	$\mathbf{76.50}_{3.14}$	**48.60**

Table 9: Classification performance (Macro F1(%)) on OPT-2.7B. mean_{std}, top-2 results are in **bold**.

OPT 2.7B		AGNews	SemE.R	SemE.L	PoemS	RTE	TEE	TEH	TES	Fina.P	Rott.T	Average
k=0	None	$27.67_{0.00}$	$24.62_{0.00}$	$31.07_{0.00}$	$25.00_{0.00}$	$51.19_{0.00}$	$24.43_{0.00}$	$39.31_{0.00}$	$34.08_{0.00}$	$\mathbf{47.23}_{0.00}$	$40.33_{0.00}$	34.49
	Con.C	$20.72_{0.00}$	$20.77_{0.00}$	$29.33_{0.00}$	$16.51_{0.00}$	$34.63_{0.00}$	$16.25_{0.00}$	$31.37_{0.00}$	$30.65_{0.00}$	$44.88_{0.00}$	$\mathbf{67.48}_{0.00}$	31.26
	Bat.C	$28.58_{0.00}$	$31.32_{0.00}$	$34.31_{0.00}$	$25.44_{0.00}$	$\mathbf{53.23}_{0.00}$	$26.81_{0.00}$	$43.60_{0.00}$	$34.34_{0.00}$	$45.25_{0.00}$	$\mathbf{66.51}_{0.00}$	38.94
	Dom.C	$27.55_{0.06}$	$20.53_{0.05}$	$30.90_{0.84}$	$16.72_{0.40}$	$45.07_{0.12}$	$18.01_{0.18}$	$43.30_{0.12}$	$31.11_{0.24}$	$\mathbf{46.16}_{0.47}$	$66.23_{1.28}$	34.56
	KNN	$52.50_{0.00}$	$31.39_{0.00}$	$\mathbf{44.16}_{0.00}$	$25.72_{0.00}$	$\mathbf{52.08}_{0.00}$	$36.15_{0.00}$	$51.60_{0.00}$	$\mathbf{39.45}_{0.00}$	$42.26_{0.00}$	$57.16_{0.00}$	43.25
	Cent.C	$\mathbf{55.97}_{0.00}$	$\mathbf{31.65}_{0.00}$	$43.23_{0.00}$	$\mathbf{35.86}_{0.00}$	$41.98_{0.00}$	$\mathbf{41.27}_{0.00}$	$\mathbf{53.44}_{0.00}$	$35.25_{0.00}$	$37.54_{0.00}$	$58.60_{0.00}$	**43.48**
	Hidd.C	$\mathbf{75.01}_{0.00}$	$\mathbf{41.94}_{0.00}$	$\mathbf{52.14}_{0.00}$	$\mathbf{39.92}_{0.00}$	$45.64_{0.00}$	$\mathbf{45.90}_{0.00}$	$\mathbf{52.93}_{0.00}$	$\mathbf{35.67}_{0.00}$	$43.24_{0.00}$	$61.71_{0.00}$	**49.41**
k=1	None	$24.16_{0.71}$	$18.76_{0.50}$	$24.98_{1.74}$	$12.23_{1.19}$	$\mathbf{50.75}_{3.01}$	$20.72_{3.33}$	$\mathbf{51.55}_{4.28}$	$23.64_{1.19}$	$31.40_{1.83}$	$48.94_{1.90}$	30.71
	Con.C	$22.57_{1.38}$	$20.15_{1.78}$	$24.48_{1.88}$	$13.17_{0.63}$	$\mathbf{50.75}_{3.01}$	$24.30_{1.69}$	$\mathbf{51.69}_{4.03}$	$23.02_{0.00}$	$28.88_{1.95}$	$65.63_{1.72}$	32.46
	Bat.C	$27.43_{1.13}$	$20.48_{0.90}$	$26.78_{1.17}$	$14.21_{1.05}$	$50.25_{3.25}$	$26.44_{1.41}$	$50.45_{1.76}$	$23.34_{2.30}$	$28.81_{0.24}$	$\mathbf{70.88}_{0.65}$	33.91
	Dom.C	$24.06_{0.94}$	$19.68_{1.26}$	$24.24_{1.73}$	$13.78_{2.27}$	$\mathbf{50.75}_{3.01}$	$23.58_{2.54}$	$51.54_{3.86}$	$23.14_{0.83}$	$28.81_{1.72}$	$\mathbf{69.98}_{1.81}$	32.96
	KNN	$48.15_{2.50}$	$\mathbf{42.35}_{4.06}$	$39.01_{5.24}$	$25.52_{1.94}$	$\mathbf{53.07}_{3.06}$	$32.62_{3.02}$	$49.78_{0.58}$	$\mathbf{31.90}_{3.64}$	$36.59_{3.67}$	$57.31_{4.89}$	**41.63**
	Cent.C	$\mathbf{49.21}_{2.81}$	$39.46_{4.26}$	$\mathbf{42.30}_{4.65}$	$\mathbf{29.60}_{5.30}$	$48.48_{3.85}$	$\mathbf{34.32}_{2.51}$	$50.69_{1.21}$	$31.00_{1.34}$	$\mathbf{36.69}_{2.28}$	$53.70_{2.51}$	41.54
	Hidd.C	$\mathbf{65.18}_{2.39}$	$\mathbf{44.91}_{5.14}$	$\mathbf{51.62}_{2.09}$	$\mathbf{33.72}_{2.24}$	$50.40_{1.80}$	$\mathbf{45.00}_{2.79}$	$49.53_{2.92}$	$\mathbf{35.02}_{0.47}$	$\mathbf{48.10}_{3.33}$	$57.72_{1.42}$	**48.12**
k=4	None	$22.91_{1.05}$	$20.84_{1.16}$	$25.44_{1.60}$	$12.46_{1.29}$	$49.70_{3.22}$	$14.83_{0.16}$	$40.68_{0.44}$	$23.78_{0.99}$	$28.62_{1.13}$	$53.73_{0.95}$	29.30
	Con.C	$22.81_{1.04}$	$20.70_{1.02}$	$27.74_{1.79}$	$12.62_{1.32}$	$\mathbf{50.06}_{3.86}$	$21.92_{1.03}$	$41.21_{0.70}$	$25.86_{1.65}$	$36.74_{1.68}$	$\mathbf{83.50}_{1.82}$	34.32
	Bat.C	$25.40_{1.04}$	$20.07_{1.40}$	$26.49_{0.86}$	$11.46_{1.24}$	$47.61_{2.78}$	$23.92_{0.96}$	$45.71_{1.33}$	$26.91_{1.24}$	$38.88_{1.70}$	$82.22_{1.13}$	34.87
	Dom.C	$22.32_{1.17}$	$20.71_{1.39}$	$25.96_{1.97}$	$12.64_{1.24}$	$\mathbf{50.37}_{3.22}$	$21.74_{1.30}$	$41.30_{0.54}$	$26.15_{1.41}$	$38.01_{1.69}$	$\mathbf{83.99}_{1.45}$	34.32
	KNN	$49.36_{2.28}$	$49.71_{5.26}$	$47.66_{3.87}$	$22.18_{1.17}$	$48.74_{3.76}$	$\mathbf{32.05}_{4.50}$	$\mathbf{47.35}_{3.18}$	$\mathbf{30.76}_{3.92}$	$\mathbf{40.78}_{1.86}$	$66.33_{4.19}$	**43.49**
	Cent.C	$\mathbf{49.91}_{4.46}$	$\mathbf{51.72}_{2.80}$	$\mathbf{48.96}_{3.00}$	$\mathbf{21.63}_{3.55}$	$47.02_{2.37}$	$29.84_{3.43}$	$46.99_{2.05}$	$25.39_{3.05}$	$40.23_{8.32}$	$65.80_{4.54}$	42.75
	Hidd.C	$\mathbf{69.56}_{3.62}$	$\mathbf{59.59}_{1.97}$	$\mathbf{62.77}_{3.08}$	$\mathbf{25.42}_{0.79}$	$47.94_{3.16}$	$\mathbf{47.47}_{5.86}$	$\mathbf{49.46}_{1.80}$	$\mathbf{35.61}_{3.45}$	$\mathbf{65.91}_{2.01}$	$79.47_{5.88}$	**54.32**
k=8	None	$21.52_{1.03}$	$21.24_{1.48}$	$24.11_{1.51}$	$12.84_{0.73}$	$48.05_{3.65}$	$14.54_{0.00}$	$37.54_{0.03}$	$21.85_{1.07}$	$28.47_{1.48}$	$59.86_{1.63}$	29.00
	Con.C	$22.26_{0.33}$	$23.84_{1.26}$	$30.27_{1.27}$	$13.02_{0.89}$	$47.58_{3.27}$	$23.06_{1.10}$	$37.80_{0.48}$	$25.91_{0.79}$	$37.67_{1.44}$	$83.42_{1.74}$	34.48
	Bat.C	$24.38_{1.49}$	$22.28_{1.04}$	$28.05_{1.65}$	$13.31_{0.50}$	$47.50_{2.86}$	$24.46_{0.50}$	$49.30_{1.63}$	$27.96_{1.51}$	$40.39_{1.53}$	$84.03_{1.31}$	36.16
	Dom.C	$22.38_{0.70}$	$21.33_{1.57}$	$26.63_{1.36}$	$13.18_{1.08}$	$47.48_{3.87}$	$23.63_{1.08}$	$37.59_{0.25}$	$25.25_{1.11}$	$39.27_{2.20}$	$\mathbf{86.67}_{1.55}$	34.34
	KNN	$47.45_{1.64}$	$55.08_{1.89}$	$48.16_{3.38}$	$22.61_{2.20}$	$\mathbf{49.56}_{2.68}$	$\mathbf{34.80}_{3.18}$	$49.60_{2.26}$	$\mathbf{30.77}_{2.79}$	$38.73_{3.52}$	$66.65_{5.67}$	44.34
	Cent.C	$\mathbf{47.50}_{4.46}$	$\mathbf{55.83}_{4.46}$	$\mathbf{51.06}_{8.42}$	$\mathbf{22.95}_{2.22}$	$48.22_{2.35}$	$33.57_{4.12}$	$\mathbf{49.66}_{2.73}$	$29.18_{4.37}$	$\mathbf{43.87}_{2.57}$	$67.49_{8.07}$	**44.93**
	Hidd.C	$\mathbf{65.27}_{1.64}$	$\mathbf{63.95}_{2.51}$	$\mathbf{66.48}_{2.53}$	$\mathbf{23.70}_{1.76}$	$\mathbf{49.86}_{2.86}$	$\mathbf{52.24}_{2.39}$	$\mathbf{50.57}_{3.28}$	$\mathbf{41.94}_{6.42}$	$\mathbf{71.16}_{4.13}$	$\mathbf{85.78}_{4.39}$	**57.09**

Table 10: Classification performance (Macro F1(%)) on LLaMa 2-6.9B. mean_{std}, top-2 results are in **bold**.

LLaMa 2		AGNews	SemE.R	SemE.L	PoemS	RTE	TEE	TEH	TES	Fina.P	Rott.T	Average
k=0	None	$23.74_{0.00}$	$44.36_{0.00}$	$32.48_{0.00}$	$16.23_{0.00}$	$35.88_{0.00}$	$17.01_{0.00}$	$46.06_{0.00}$	$30.59_{0.00}$	$19.64_{0.00}$	$32.90_{0.00}$	29.89
	Con.C	$22.76_{0.00}$	$44.50_{0.00}$	$42.83_{0.00}$	$10.53_{0.00}$	$43.88_{0.00}$	$24.06_{0.00}$	$42.42_{0.00}$	$33.32_{0.00}$	$54.92_{0.00}$	$\mathbf{58.97}_{0.00}$	37.82
	Bat.C	$37.26_{0.00}$	$49.03_{0.00}$	$\mathbf{54.33}_{0.00}$	$19.70_{0.00}$	$\mathbf{47.34}_{0.00}$	$\mathbf{31.64}_{0.00}$	$\mathbf{51.06}_{0.00}$	$\mathbf{39.82}_{0.00}$	$\mathbf{55.76}_{0.00}$	$\mathbf{63.07}_{0.00}$	**44.90**
	Dom.C	$28.11_{1.43}$	$47.88_{0.45}$	$49.74_{1.40}$	$\mathbf{26.58}_{1.27}$	$39.99_{0.13}$	$23.00_{1.11}$	$37.48_{0.00}$	$33.33_{0.21}$	$\mathbf{61.03}_{0.76}$	$54.96_{2.78}$	40.21
	KNN	$\mathbf{40.24}_{0.00}$	$47.65_{0.00}$	$49.15_{0.00}$	$20.56_{0.00}$	$\mathbf{50.93}_{0.00}$	$26.65_{0.00}$	$42.35_{0.00}$	$33.63_{0.00}$	$46.44_{0.00}$	$53.26_{0.00}$	41.09
	Cent.C	$34.36_{0.00}$	$\mathbf{50.87}_{0.00}$	$46.36_{0.00}$	$17.57_{0.00}$	$42.15_{0.00}$	$25.09_{0.00}$	$42.31_{0.00}$	$28.21_{0.00}$	$43.81_{0.00}$	$49.38_{0.00}$	38.01
	Hidd.C	$\mathbf{62.46}_{0.00}$	$\mathbf{50.90}_{0.00}$	$\mathbf{55.77}_{0.00}$	$\mathbf{22.49}_{0.00}$	$46.94_{0.00}$	$\mathbf{34.06}_{0.00}$	$\mathbf{47.37}_{0.00}$	$\mathbf{34.29}_{0.00}$	$54.84_{0.00}$	$56.37_{0.00}$	**46.49**
k=1	None	$13.97_{1.15}$	$51.57_{0.30}$	$51.50_{0.47}$	$22.28_{0.98}$	$36.26_{0.80}$	$27.80_{1.58}$	$40.16_{0.83}$	$30.78_{2.77}$	$27.89_{2.18}$	$55.03_{1.91}$	35.72
	Con.C	$13.29_{1.62}$	$52.36_{0.58}$	$54.12_{0.50}$	$24.16_{1.29}$	$38.53_{0.92}$	$27.77_{1.16}$	$40.63_{1.52}$	$37.39_{1.86}$	$37.23_{0.68}$	$\mathbf{75.91}_{2.44}$	40.14
	Bat.C	$22.05_{1.07}$	$52.24_{0.44}$	$53.77_{1.27}$	$\mathbf{24.75}_{1.51}$	$\mathbf{49.77}_{3.25}$	$28.94_{0.90}$	$47.72_{1.19}$	$\mathbf{40.25}_{2.30}$	$38.43_{1.91}$	$75.77_{1.37}$	43.37
	Dom.C	$11.89_{1.01}$	$51.80_{0.38}$	$53.98_{0.45}$	$24.23_{1.13}$	$\mathbf{49.33}_{3.82}$	$29.17_{1.44}$	$38.23_{0.85}$	$39.02_{1.72}$	$34.95_{1.35}$	$\mathbf{79.44}_{0.58}$	41.20
	KNN	$\mathbf{52.50}_{3.58}$	$56.19_{5.59}$	$61.24_{2.42}$	$21.91_{0.91}$	$48.82_{1.90}$	$\mathbf{30.66}_{3.56}$	$\mathbf{49.65}_{3.15}$	$34.86_{1.11}$	$\mathbf{40.26}_{3.40}$	$75.26_{3.34}$	**46.87**
	Cent.C	$46.17_{0.93}$	$\mathbf{62.09}_{3.04}$	$\mathbf{64.50}_{1.76}$	$23.29_{2.57}$	$46.66_{3.16}$	$27.80_{2.36}$	$45.63_{2.33}$	$33.10_{2.42}$	$38.99_{3.56}$	$75.16_{2.33}$	46.34
	Hidd.C	$\mathbf{61.88}_{0.68}$	$\mathbf{64.83}_{1.85}$	$\mathbf{69.05}_{2.29}$	$\mathbf{24.26}_{1.37}$	$48.32_{2.35}$	$\mathbf{42.77}_{2.38}$	$\mathbf{51.55}_{1.27}$	$\mathbf{40.72}_{1.91}$	$\mathbf{63.79}_{0.93}$	$71.00_{5.23}$	**53.82**
k=4	None	$9.57_{0.26}$	$53.45_{0.34}$	$53.76_{0.62}$	$19.82_{0.51}$	$35.90_{0.18}$	$29.30_{1.47}$	$37.56_{0.00}$	$36.55_{0.97}$	$37.86_{0.78}$	$69.31_{1.43}$	38.31
	Con.C	$9.35_{0.23}$	$53.49_{0.33}$	$54.21_{0.30}$	$\mathbf{27.01}_{1.15}$	$37.65_{1.74}$	$\mathbf{31.37}_{1.07}$	$37.56_{0.00}$	$39.34_{0.47}$	$39.28_{0.55}$	$90.86_{0.35}$	42.01
	Bat.C	$19.33_{1.51}$	$52.58_{0.64}$	$54.40_{0.69}$	$24.62_{0.99}$	$46.81_{5.22}$	$31.26_{1.58}$	$48.81_{2.93}$	$38.47_{0.33}$	$37.05_{0.60}$	$86.62_{1.15}$	44.00
	Dom.C	$9.49_{0.32}$	$52.12_{0.79}$	$54.14_{0.28}$	$\mathbf{26.47}_{0.87}$	$39.59_{2.03}$	$30.75_{1.06}$	$37.56_{1.43}$	$\mathbf{39.52}_{0.22}$	$34.71_{0.58}$	$\mathbf{91.84}_{0.88}$	41.62
	KNN	$\mathbf{66.47}_{6.82}$	$60.00_{0.92}$	$62.87_{2.54}$	$21.76_{3.20}$	$54.41_{4.95}$	$29.96_{2.19}$	$49.46_{2.22}$	$31.90_{3.97}$	$\mathbf{49.09}_{5.20}$	$88.18_{1.17}$	**51.41**
	Cent.C	$60.62_{4.62}$	$\mathbf{63.61}_{2.06}$	$\mathbf{64.55}_{3.77}$	$21.42_{4.73}$	$\mathbf{55.50}_{5.61}$	$28.41_{2.51}$	$\mathbf{50.72}_{3.29}$	$21.12_{5.70}$	$42.48_{2.81}$	$89.55_{0.74}$	49.80
	Hidd.C	$\mathbf{68.22}_{6.82}$	$\mathbf{65.64}_{1.38}$	$\mathbf{71.38}_{2.00}$	$25.47_{2.10}$	$\mathbf{58.00}_{4.38}$	$\mathbf{48.48}_{4.32}$	$\mathbf{51.83}_{3.03}$	$\mathbf{49.44}_{3.20}$	$\mathbf{68.79}_{9.71}$	$\mathbf{92.26}_{0.18}$	**59.95**
k=8	None	$8.03_{0.00}$	$53.90_{0.31}$	$54.12_{0.38}$	$20.47_{0.49}$	$35.98_{0.00}$	$30.77_{1.50}$	$37.56_{0.00}$	$38.59_{1.26}$	$39.25_{1.04}$	$73.77_{0.60}$	39.24
	Con.C	$8.15_{0.26}$	$53.72_{0.30}$	$54.26_{0.37}$	$\mathbf{28.72}_{1.91}$	$35.98_{0.00}$	$31.40_{1.94}$	$37.56_{0.00}$	$39.33_{0.53}$	$38.80_{0.46}$	$91.52_{0.38}$	41.94
	Bat.C	$17.58_{1.19}$	$53.06_{0.52}$	$54.15_{0.28}$	$25.62_{0.85}$	$51.85_{3.70}$	$30.66_{1.48}$	$48.45_{2.63}$	$38.44_{0.46}$	$36.28_{0.37}$	$88.11_{0.62}$	44.42
	Dom.C	$8.03_{0.00}$	$52.26_{0.43}$	$54.26_{0.47}$	$26.50_{0.99}$	$36.44_{1.05}$	$29.28_{2.06}$	$37.56_{0.00}$	$\mathbf{39.72}_{0.41}$	$35.21_{0.86}$	$\mathbf{91.91}_{0.49}$	41.12
	KNN	$\mathbf{69.41}_{4.38}$	$59.12_{4.16}$	$62.69_{3.74}$	$23.15_{2.09}$	$\mathbf{54.86}_{4.00}$	$\mathbf{34.89}_{2.06}$	$48.44_{3.62}$	$36.27_{4.36}$	$\mathbf{50.27}_{5.38}$	$89.60_{1.32}$	**52.87**
	Cent.C	$\mathbf{60.04}_{8.07}$	$\mathbf{61.80}_{1.44}$	$\mathbf{66.16}_{2.05}$	$23.17_{1.31}$	$53.64_{3.23}$	$32.16_{2.62}$	$\mathbf{51.74}_{1.94}$	$30.29_{3.91}$	$36.55_{6.71}$	$90.12_{0.52}$	50.57
	Hidd.C	$59.17_{4.08}$	$\mathbf{65.88}_{0.84}$	$\mathbf{70.07}_{2.32}$	$\mathbf{27.85}_{2.56}$	$\mathbf{55.79}_{3.05}$	$\mathbf{46.70}_{5.37}$	$\mathbf{53.95}_{2.36}$	$\mathbf{43.90}_{0.48}$	$\mathbf{67.69}_{2.29}$	$\mathbf{92.26}_{0.63}$	**58.33**

Table 11: Classification performance (Macro F1(%)) on LLaMa 3-8B. mean_{std}, top-2 results are in **bold**.

LLaMa 3		AGNews	SemE.R	SemE.L	PoemS	RTE	TEE	TEH	TES	Fina.P	Rott.T	Average
k=0	None	$19.55_{0.00}$	$27.61_{0.00}$	$24.38_{0.00}$	$11.25_{0.00}$	$43.68_{0.00}$	$26.33_{0.00}$	$46.69_{0.00}$	$30.67_{0.00}$	$44.45_{0.00}$	$61.19_{0.00}$	33.58
	Con.C	$23.20_{0.00}$	$17.16_{0.00}$	$14.55_{0.00}$	$2.33_{0.00}$	$\mathbf{47.50}_{0.00}$	$21.66_{0.00}$	$33.34_{0.00}$	$29.66_{0.00}$	$28.78_{0.00}$	$60.71_{0.00}$	27.89
	Bat.C	$26.59_{0.00}$	$43.52_{0.00}$	$38.62_{0.00}$	$17.40_{0.00}$	$47.25_{0.00}$	$\mathbf{32.02}_{0.00}$	$\mathbf{49.12}_{0.00}$	$\mathbf{38.17}_{0.00}$	$\mathbf{50.32}_{0.00}$	$\mathbf{67.19}_{0.00}$	41.02
	Dom.C	$19.17_{0.00}$	$25.92_{0.00}$	$24.36_{0.00}$	$3.97_{0.00}$	$32.86_{0.00}$	$26.46_{0.00}$	$39.35_{0.00}$	$\mathbf{37.95}_{0.00}$	$\mathbf{48.39}_{0.00}$	$59.50_{0.00}$	31.79
	KNN	$\mathbf{49.70}_{0.00}$	$51.20_{0.00}$	$47.55_{0.00}$	$20.74_{0.00}$	$47.25_{0.00}$	$28.97_{0.00}$	$51.08_{0.00}$	$31.26_{0.00}$	$47.80_{0.00}$	$52.64_{0.00}$	**42.82**
	Cent.C	$39.40_{0.00}$	$\mathbf{53.41}_{0.00}$	$\mathbf{48.67}_{0.00}$	$\mathbf{23.68}_{0.00}$	$\mathbf{47.41}_{0.00}$	$27.00_{0.00}$	$\mathbf{49.58}_{0.00}$	$25.19_{0.00}$	$40.38_{0.00}$	$51.14_{0.00}$	40.58
	Hidd.C	$\mathbf{81.52}_{0.00}$	$\mathbf{52.75}_{0.00}$	$\mathbf{49.43}_{0.00}$	$\mathbf{27.67}_{0.00}$	$44.53_{0.00}$	$\mathbf{34.04}_{0.00}$	$44.53_{0.00}$	$37.20_{0.00}$	$47.74_{0.00}$	$\mathbf{63.04}_{0.00}$	**48.25**
k=1	None	$19.97_{0.70}$	$53.47_{0.13}$	$53.75_{0.74}$	$\mathbf{25.41}_{1.06}$	$48.60_{2.81}$	$17.02_{1.72}$	$\mathbf{50.68}_{2.19}$	$28.30_{1.50}$	$39.28_{1.36}$	$81.04_{2.00}$	41.75
	Con.C	$15.57_{1.02}$	$53.37_{0.69}$	$54.42_{0.44}$	$20.37_{5.23}$	$48.60_{3.06}$	$19.27_{1.28}$	$49.71_{3.07}$	$37.32_{1.11}$	$39.07_{0.33}$	$\mathbf{90.32}_{0.74}$	42.80
	Bat.C	$27.50_{1.91}$	$52.87_{0.17}$	$55.71_{0.60}$	$23.73_{0.65}$	$49.23_{2.95}$	$26.48_{0.99}$	$49.80_{2.77}$	$\mathbf{39.54}_{2.26}$	$39.06_{0.76}$	$87.01_{0.79}$	45.09
	Dom.C	$12.96_{1.62}$	$52.23_{1.07}$	$54.41_{0.56}$	$19.94_{4.35}$	$\mathbf{49.37}_{3.03}$	$20.24_{0.86}$	$49.73_{3.05}$	$37.86_{1.05}$	$36.40_{0.36}$	$\mathbf{90.27}_{0.56}$	42.34
	KNN	$\mathbf{61.64}_{0.94}$	$54.58_{3.37}$	$61.94_{1.25}$	$21.49_{1.40}$	$48.15_{2.60}$	$34.73_{2.96}$	$48.13_{2.61}$	$39.11_{1.54}$	$\mathbf{43.88}_{1.85}$	$87.67_{1.78}$	**50.13**
	Cent.C	$60.16_{4.68}$	$\mathbf{56.81}_{5.45}$	$\mathbf{62.72}_{1.87}$	$22.74_{3.75}$	$45.77_{4.80}$	$\mathbf{35.10}_{3.40}$	$44.59_{5.76}$	$35.95_{2.01}$	$38.15_{2.52}$	$88.06_{1.14}$	49.00
	Hidd.C	$\mathbf{83.37}_{2.76}$	$\mathbf{60.59}_{5.42}$	$\mathbf{64.12}_{1.73}$	$\mathbf{27.82}_{1.99}$	$\mathbf{49.24}_{1.88}$	$\mathbf{53.37}_{3.10}$	$\mathbf{51.77}_{2.59}$	$\mathbf{41.87}_{1.07}$	$\mathbf{65.62}_{5.94}$	$90.17_{0.58}$	**58.79**
k=4	None	$9.75_{0.22}$	$53.90_{0.25}$	$54.36_{0.52}$	$\mathbf{24.45}_{1.23}$	$37.09_{1.14}$	$14.54_{0.00}$	$39.42_{1.76}$	$33.23_{1.55}$	$40.28_{0.18}$	$84.80_{0.93}$	39.18
	Con.C	$9.17_{0.00}$	$53.72_{0.64}$	$54.42_{0.51}$	$\mathbf{23.58}_{4.38}$	$37.20_{1.21}$	$15.56_{0.72}$	$43.43_{4.03}$	$38.31_{1.40}$	$40.17_{1.84}$	$\mathbf{92.71}_{0.63}$	40.83
	Bat.C	$23.64_{0.50}$	$52.70_{0.67}$	$54.35_{0.33}$	$22.95_{1.60}$	$46.73_{5.45}$	$27.62_{1.01}$	$\mathbf{51.46}_{0.44}$	$\mathbf{38.65}_{0.99}$	$37.11_{0.41}$	$90.51_{1.13}$	44.57
	Dom.C	$9.17_{0.01}$	$52.51_{0.47}$	$54.35_{0.38}$	$21.85_{1.37}$	$37.52_{1.69}$	$15.94_{0.73}$	$43.01_{5.37}$	$38.04_{1.01}$	$37.76_{0.82}$	$\mathbf{92.57}_{0.48}$	40.27
	KNN	$\mathbf{74.94}_{3.56}$	$\mathbf{54.73}_{5.26}$	$\mathbf{59.53}_{2.92}$	$23.01_{3.28}$	$\mathbf{49.54}_{5.90}$	$\mathbf{40.01}_{1.41}$	$49.43_{3.14}$	$37.00_{4.50}$	$42.38_{2.07}$	$86.28_{1.79}$	**51.69**
	Cent.C	$73.26_{3.26}$	$54.31_{3.12}$	$61.26_{4.08}$	$21.02_{2.43}$	$45.64_{10.22}$	$38.73_{5.55}$	$50.74_{5.16}$	$37.05_{3.14}$	$\mathbf{44.61}_{2.96}$	$85.45_{2.18}$	51.21
	Hidd.C	$\mathbf{84.53}_{1.35}$	$\mathbf{61.73}_{4.56}$	$\mathbf{69.28}_{1.43}$	$23.03_{1.82}$	$\mathbf{53.49}_{3.61}$	$\mathbf{54.70}_{2.03}$	$\mathbf{52.96}_{3.84}$	$\mathbf{40.71}_{2.35}$	$\mathbf{72.18}_{3.19}$	$92.18_{1.39}$	**60.48**
k=8	None	-	$53.77_{0.42}$	$54.82_{0.12}$	$\mathbf{27.10}_{1.62}$	$35.98_{0.00}$	$14.54_{0.00}$	$37.69_{0.27}$	$35.96_{1.51}$	$42.65_{0.85}$	$88.40_{0.59}$	39.09
	Con.C	-	$53.52_{0.53}$	$54.42_{0.53}$	$\mathbf{25.07}_{4.81}$	$35.98_{0.00}$	$14.54_{0.00}$	$47.99_{9.84}$	$38.74_{1.35}$	$41.19_{2.12}$	$\mathbf{92.81}_{1.09}$	40.43
	Bat.C	-	$52.40_{0.72}$	$54.82_{0.35}$	$24.68_{0.67}$	$\mathbf{60.38}_{5.04}$	$31.50_{0.44}$	$\mathbf{55.00}_{0.54}$	$38.59_{0.78}$	$37.95_{0.35}$	$91.59_{0.53}$	44.69
	Dom.C	-	$52.03_{0.35}$	$54.11_{0.66}$	$21.95_{0.83}$	$35.98_{0.00}$	$14.67_{0.22}$	$45.15_{9.11}$	$38.61_{1.13}$	$37.30_{1.39}$	$93.11_{0.74}$	39.29
	KNN	-	$54.47_{2.60}$	$\mathbf{58.17}_{2.84}$	$21.49_{1.69}$	$50.75_{4.92}$	$\mathbf{47.22}_{2.50}$	$51.75_{2.65}$	$\mathbf{41.66}_{3.19}$	$\mathbf{46.47}_{4.53}$	$84.53_{1.91}$	**45.65**
	Cent.C	-	$\mathbf{56.03}_{3.55}$	$49.78_{5.61}$	$22.15_{1.36}$	$48.30_{9.65}$	$45.41_{6.52}$	$48.34_{4.57}$	$38.65_{1.60}$	$39.58_{7.11}$	$85.23_{2.05}$	43.35
	Hidd.C	-	$\mathbf{63.34}_{1.61}$	$\mathbf{67.48}_{1.00}$	$23.84_{1.22}$	$\mathbf{60.89}_{3.38}$	$\mathbf{56.67}_{7.36}$	$\mathbf{53.45}_{4.92}$	$\mathbf{39.74}_{2.66}$	$\mathbf{79.28}_{4.45}$	$\mathbf{91.64}_{1.59}$	**53.63**

Table 12: Classification performance (Macro F1(%)) on LLaMa 2-13B. mean_{std}, top-2 results are in **bold**.

LLaMa 2		AGNews	SemE.R	SemE.L	PoemS	RTE	TEE	TEH	TES	Fina.P	Rott.T	Average
k=0	None	$9.61_{0.00}$	$47.67_{0.00}$	$51.12_{0.00}$	$8.01_{0.00}$	$36.04_{0.00}$	$31.92_{0.00}$	$37.56_{0.00}$	$46.19_{0.00}$	$30.94_{0.00}$	$54.98_{0.00}$	35.40
	Con.C	$22.36_{0.00}$	$48.44_{0.00}$	$\mathbf{57.96}_{0.00}$	$15.71_{0.00}$	$41.45_{0.00}$	$23.57_{0.00}$	$37.56_{0.00}$	$39.59_{0.00}$	$44.97_{0.00}$	$42.61_{0.00}$	37.42
	Bat.C	$30.84_{0.00}$	$\mathbf{52.16}_{0.00}$	$\mathbf{61.45}_{0.00}$	$20.50_{0.00}$	$43.54_{0.00}$	$\mathbf{37.88}_{0.00}$	$43.73_{0.00}$	$\mathbf{47.31}_{0.00}$	$\mathbf{50.83}_{0.00}$	$\mathbf{61.26}_{0.00}$	**44.95**
	Dom.C	$23.21_{0.00}$	$46.54_{0.00}$	$51.51_{0.00}$	$8.88_{0.00}$	$41.87_{0.00}$	$29.11_{0.00}$	$37.56_{0.00}$	$\mathbf{47.22}_{0.00}$	$31.98_{0.00}$	$\mathbf{60.59}_{0.00}$	37.85
	KNN	$\mathbf{46.44}_{0.00}$	$42.06_{0.00}$	$43.18_{0.00}$	$\mathbf{24.03}_{0.00}$	$\mathbf{50.33}_{0.00}$	$28.36_{0.00}$	$\mathbf{55.76}_{0.00}$	$35.27_{0.00}$	$45.74_{0.00}$	$51.95_{0.00}$	42.31
	Cent.C	$43.48_{0.00}$	$44.64_{0.00}$	$41.05_{0.00}$	$21.23_{0.00}$	$34.02_{0.00}$	$30.41_{0.00}$	$49.02_{0.00}$	$30.44_{0.00}$	$43.17_{0.00}$	$52.75_{0.00}$	39.02
	Hidd.C	$\mathbf{69.78}_{0.00}$	$\mathbf{54.08}_{0.00}$	$50.67_{0.00}$	$\mathbf{33.01}_{0.00}$	$\mathbf{54.41}_{0.00}$	$\mathbf{41.23}_{0.00}$	$\mathbf{51.14}_{0.00}$	$35.45_{0.00}$	$\mathbf{48.87}_{0.00}$	$57.61_{0.00}$	**49.63**
k=1	None	$9.17_{0.00}$	$50.67_{0.80}$	$53.51_{0.08}$	$15.49_{0.67}$	$35.11_{0.00}$	$33.79_{1.09}$	$37.56_{0.00}$	$37.11_{0.78}$	$34.73_{0.90}$	$81.61_{0.84}$	38.88
	Con.C	$9.17_{0.00}$	$53.38_{0.30}$	$\mathbf{54.04}_{0.20}$	$\mathbf{25.20}_{0.64}$	$35.11_{0.00}$	$33.61_{0.46}$	$37.52_{0.05}$	$\mathbf{38.00}_{0.53}$	$35.73_{0.58}$	$\mathbf{87.40}_{1.46}$	40.92
	Bat.C	$24.81_{2.72}$	$49.21_{0.87}$	$53.72_{0.14}$	$23.39_{0.33}$	$47.24_{1.16}$	$\mathbf{35.68}_{0.19}$	$\mathbf{51.66}_{1.75}$	$36.21_{0.81}$	$34.35_{0.39}$	$85.99_{1.19}$	44.23
	Dom.C	$9.55_{0.38}$	$51.02_{0.55}$	$52.92_{0.29}$	$19.66_{1.43}$	$35.11_{0.00}$	$32.55_{2.39}$	$37.56_{0.00}$	$38.43_{0.72}$	$35.55_{0.68}$	$\mathbf{86.42}_{1.75}$	39.88
	KNN	$\mathbf{58.35}_{3.83}$	$\mathbf{54.98}_{5.17}$	$48.86_{0.44}$	$24.03_{1.43}$	$\mathbf{47.50}_{2.83}$	$32.11_{1.50}$	$50.95_{3.28}$	$33.33_{4.95}$	$\mathbf{37.93}_{3.72}$	$82.19_{1.11}$	**47.02**
	Cent.C	$51.96_{0.88}$	$49.83_{4.34}$	$46.51_{2.10}$	$25.20_{2.94}$	$43.15_{3.01}$	$27.11_{1.80}$	$47.81_{6.69}$	$31.46_{6.15}$	$31.22_{4.39}$	$82.31_{1.10}$	43.65
	Hidd.C	$\mathbf{76.45}_{5.43}$	$\mathbf{64.70}_{1.97}$	$\mathbf{65.70}_{0.16}$	$\mathbf{30.22}_{3.11}$	$\mathbf{49.42}_{1.78}$	$\mathbf{48.97}_{0.32}$	$\mathbf{52.81}_{4.66}$	$\mathbf{41.05}_{4.32}$	$\mathbf{59.29}_{4.43}$	$84.02_{0.88}$	**57.26**
k=4	None	$9.17_{0.00}$	$49.99_{0.32}$	$53.69_{0.44}$	$19.29_{0.86}$	$35.11_{0.00}$	$\mathbf{36.23}_{2.22}$	$37.56_{0.00}$	$37.63_{0.58}$	$36.02_{0.42}$	$91.82_{0.41}$	40.65
	Con.C	$9.37_{0.22}$	$53.12_{0.62}$	$\mathbf{54.25}_{0.26}$	$\mathbf{26.98}_{1.34}$	$35.11_{0.00}$	$34.64_{2.69}$	$37.56_{0.00}$	$37.64_{1.28}$	$36.38_{0.15}$	$\mathbf{92.52}_{0.76}$	41.76
	Bat.C	$19.64_{1.36}$	$49.33_{0.12}$	$53.69_{0.13}$	$23.93_{0.41}$	$46.56_{3.27}$	$34.29_{1.84}$	$46.95_{1.79}$	$35.92_{0.83}$	$34.14_{0.46}$	$91.68_{0.52}$	43.61
	Dom.C	$9.56_{0.31}$	$49.61_{0.38}$	$53.79_{0.35}$	$16.85_{0.80}$	$35.11_{0.00}$	$33.62_{1.59}$	$37.56_{0.00}$	$\mathbf{38.20}_{1.38}$	$35.95_{0.35}$	$\mathbf{92.46}_{0.83}$	40.27
	KNN	$\mathbf{76.62}_{0.69}$	$\mathbf{53.18}_{1.01}$	$53.38_{4.35}$	$22.73_{1.17}$	$47.94_{1.65}$	$35.09_{5.00}$	$\mathbf{50.17}_{1.01}$	$38.26_{1.25}$	$40.40_{2.69}$	$83.18_{3.32}$	**50.09**
	Cent.C	$74.72_{7.17}$	$52.52_{2.90}$	$48.20_{3.83}$	$22.08_{2.73}$	$\mathbf{48.59}_{2.55}$	$28.15_{3.24}$	$49.53_{1.53}$	$29.35_{4.83}$	$\mathbf{43.69}_{5.11}$	$82.73_{1.48}$	47.96
	Hidd.C	$\mathbf{80.76}_{4.75}$	$\mathbf{66.91}_{1.22}$	$\mathbf{68.44}_{2.54}$	$\mathbf{25.39}_{1.27}$	$\mathbf{52.69}_{3.96}$	$\mathbf{53.41}_{3.40}$	$\mathbf{52.29}_{4.93}$	$\mathbf{44.47}_{4.05}$	$\mathbf{59.48}_{1.42}$	$91.36_{0.79}$	**59.52**
k=8	None	$9.17_{0.00}$	$50.01_{0.25}$	$53.82_{0.37}$	$19.54_{0.37}$	$35.11_{0.00}$	$36.93_{1.76}$	$37.56_{0.00}$	$36.64_{0.75}$	$36.49_{0.04}$	$\mathbf{92.80}_{0.63}$	40.81
	Con.C	$9.17_{0.00}$	$\mathbf{52.89}_{0.18}$	$\mathbf{54.31}_{0.29}$	$\mathbf{29.11}_{0.49}$	$35.11_{0.00}$	$36.44_{1.99}$	$37.56_{0.00}$	$38.24_{0.79}$	$38.10_{0.70}$	$92.72_{0.85}$	42.36
	Bat.C	$17.94_{0.26}$	$48.95_{0.69}$	$53.64_{0.45}$	$24.35_{0.44}$	$\mathbf{52.35}_{3.27}$	$34.94_{2.46}$	$47.18_{3.05}$	$36.11_{0.40}$	$35.04_{0.09}$	$92.23_{0.64}$	44.27
	Dom.C	$9.17_{0.00}$	$50.55_{0.37}$	$53.95_{0.41}$	$18.12_{1.11}$	$35.11_{0.00}$	$34.14_{2.53}$	$37.56_{0.00}$	$38.43_{0.17}$	$35.70_{0.27}$	$\mathbf{92.91}_{0.56}$	40.56
	KNN	$\mathbf{79.57}_{4.10}$	$52.03_{2.71}$	$51.68_{8.67}$	$23.48_{0.62}$	$50.35_{3.07}$	$\mathbf{39.90}_{6.24}$	$\mathbf{53.54}_{4.10}$	$\mathbf{41.95}_{2.23}$	$50.11_{3.56}$	$86.47_{1.81}$	**52.91**
	Cent.C	$76.52_{2.63}$	$51.46_{5.58}$	$42.86_{5.16}$	$22.13_{4.60}$	$49.28_{0.85}$	$34.97_{5.39}$	$53.09_{4.32}$	$33.61_{7.13}$	$\mathbf{50.29}_{2.71}$	$84.48_{2.15}$	49.87
	Hidd.C	$\mathbf{80.47}_{2.80}$	$\mathbf{66.49}_{2.19}$	$\mathbf{62.69}_{8.43}$	$\mathbf{25.58}_{3.44}$	$\mathbf{53.02}_{4.91}$	$\mathbf{50.88}_{5.15}$	$\mathbf{57.19}_{7.28}$	$\mathbf{41.99}_{3.69}$	$\mathbf{69.04}_{5.48}$	$91.31_{0.25}$	**59.87**

Table 13: Classification performance (Macro F1(%)) on LLaMa 2-34B. mean_{std}, top-2 results are in **bold**.

LLaMa 2		AGNews	SemE.R	SemE.L	PoemS	RTE	TEE	TEH	TES	Fina.P	Rott.T	Average
k=0	None	$18.28_{0.00}$	$50.09_{0.00}$	$51.30_{0.00}$	$15.50_{0.00}$	$39.75_{0.00}$	$24.91_{0.00}$	$40.14_{0.00}$	$31.05_{0.00}$	$34.67_{0.00}$	$\mathbf{71.67}_{0.00}$	37.73
	Con.C	$23.06_{0.00}$	$50.75_{0.00}$	$\mathbf{58.71}_{0.00}$	$8.81_{0.00}$	$35.11_{0.00}$	$30.96_{0.00}$	$44.55_{0.00}$	$34.93_{0.00}$	$48.68_{0.00}$	$69.32_{0.00}$	40.49
	Bat.C	$26.52_{0.00}$	$\mathbf{54.51}_{0.00}$	$\mathbf{60.19}_{0.00}$	$18.91_{0.00}$	$48.47_{0.00}$	$31.62_{0.00}$	$45.90_{0.00}$	$\mathbf{40.37}_{0.00}$	$\mathbf{53.45}_{0.00}$	$70.69_{0.00}$	**45.06**
	Dom.C	$22.58_{0.00}$	$52.91_{0.00}$	$55.96_{0.00}$	$24.06_{0.00}$	$47.61_{0.00}$	$31.18_{0.00}$	$37.30_{0.00}$	$36.45_{0.00}$	$50.42_{0.00}$	$\mathbf{71.77}_{0.00}$	43.02
	KNN	$31.79_{0.00}$	$45.85_{0.00}$	$45.03_{0.00}$	$\mathbf{27.38}_{0.00}$	$\mathbf{53.04}_{0.00}$	$31.64_{0.00}$	$\mathbf{55.29}_{0.00}$	$34.10_{0.00}$	$46.57_{0.00}$	$57.60_{0.00}$	**42.83**
	Cent.C	$\mathbf{35.28}_{0.00}$	$50.83_{0.00}$	$51.72_{0.00}$	$23.44_{0.00}$	$51.15_{0.00}$	$\mathbf{32.02}_{0.00}$	$53.12_{0.00}$	$35.33_{0.00}$	$44.25_{0.00}$	$48.47_{0.00}$	42.56
	Hidd.C	$\mathbf{65.31}_{0.00}$	$\mathbf{56.22}_{0.00}$	$54.83_{0.00}$	$\mathbf{31.82}_{0.00}$	$\mathbf{54.76}_{0.00}$	$\mathbf{50.49}_{0.00}$	$\mathbf{65.71}_{0.00}$	$\mathbf{45.46}_{0.00}$	$\mathbf{69.22}_{0.00}$	$65.14_{0.00}$	**55.90**
k=1	None	$9.59_{0.03}$	$53.48_{0.49}$	$54.97_{0.04}$	$22.51_{0.39}$	$35.53_{0.18}$	$\mathbf{33.32}_{1.18}$	$37.56_{0.38}$	$36.38_{1.24}$	$35.41_{0.18}$	$81.78_{0.31}$	40.05
	Con.C	$10.38_{0.91}$	$53.97_{0.80}$	$54.97_{0.55}$	$25.44_{0.20}$	$35.11_{2.42}$	$32.92_{0.56}$	$37.56_{0.00}$	$37.08_{2.18}$	$38.73_{0.28}$	$\mathbf{85.60}_{0.25}$	41.17
	Bat.C	$21.41_{0.28}$	$51.86_{0.43}$	$54.81_{0.10}$	$23.47_{1.50}$	$52.97_{1.18}$	$33.02_{1.00}$	$48.06_{1.46}$	$37.07_{0.67}$	$35.23_{0.21}$	$84.63_{0.44}$	44.25
	Dom.C	$10.00_{0.29}$	$53.46_{0.90}$	$54.96_{0.11}$	$\mathbf{29.23}_{2.93}$	$37.32_{0.40}$	$32.70_{0.82}$	$37.56_{0.00}$	$37.21_{1.47}$	$39.96_{0.16}$	$\mathbf{85.80}_{1.51}$	41.82
	KNN	$\mathbf{65.38}_{1.25}$	$\mathbf{58.91}_{1.13}$	$\mathbf{61.28}_{2.68}$	$29.09_{4.68}$	$\mathbf{62.63}_{3.40}$	$30.21_{1.64}$	$54.69_{0.85}$	$\mathbf{40.78}_{2.03}$	$\mathbf{68.78}_{1.13}$	$68.77_{3.17}$	**54.05**
	Cent.C	$63.03_{3.16}$	$55.45_{2.61}$	$54.41_{4.82}$	$24.23_{1.29}$	$62.47_{0.04}$	$31.16_{0.39}$	$\mathbf{56.13}_{2.05}$	$30.55_{3.73}$	$55.08_{3.23}$	$58.61_{3.26}$	49.11
	Hidd.C	$\mathbf{79.39}_{0.48}$	$\mathbf{67.23}_{2.91}$	$\mathbf{69.03}_{0.53}$	$\mathbf{33.12}_{0.35}$	$\mathbf{74.89}_{1.33}$	$\mathbf{52.50}_{4.09}$	$\mathbf{61.88}_{0.91}$	$\mathbf{46.63}_{1.04}$	$\mathbf{76.83}_{2.01}$	$73.14_{0.06}$	**63.46**
k=4	None	$9.59_{0.55}$	$53.48_{0.39}$	$54.97_{0.64}$	$22.51_{0.69}$	$35.53_{0.06}$	$\mathbf{33.32}_{3.35}$	$37.56_{0.00}$	$36.38_{1.92}$	$35.41_{0.51}$	$81.78_{0.55}$	40.05
	Con.C	$10.38_{0.01}$	$53.97_{0.16}$	$54.97_{0.61}$	$25.44_{0.06}$	$35.11_{0.00}$	$32.92_{3.78}$	$37.56_{0.00}$	$37.08_{0.97}$	$38.73_{0.55}$	$\mathbf{85.60}_{2.37}$	41.17
	Bat.C	$21.41_{1.08}$	$51.86_{0.34}$	$54.81_{0.44}$	$23.47_{0.33}$	$52.97_{0.10}$	$33.02_{4.05}$	$48.06_{2.65}$	$37.07_{1.89}$	$35.23_{0.77}$	$84.63_{1.97}$	44.25
	Dom.C	$10.00_{0.55}$	$53.46_{0.36}$	$54.96_{0.66}$	$\mathbf{29.23}_{0.90}$	$37.32_{1.28}$	$32.70_{4.18}$	$37.56_{0.00}$	$37.21_{0.80}$	$39.96_{0.04}$	$\mathbf{85.80}_{2.09}$	41.82
	KNN	$\mathbf{65.38}_{3.35}$	$\mathbf{58.91}_{0.83}$	$\mathbf{61.28}_{0.90}$	$29.09_{2.60}$	$\mathbf{62.63}_{2.71}$	$30.21_{3.04}$	$54.69_{2.18}$	$\mathbf{40.78}_{2.16}$	$\mathbf{68.78}_{1.44}$	$68.77_{8.22}$	**54.05**
	Cent.C	$63.03_{10.02}$	$55.45_{0.33}$	$54.41_{2.47}$	$24.23_{0.35}$	$62.47_{6.66}$	$31.16_{2.87}$	$\mathbf{56.13}_{0.93}$	$30.55_{0.73}$	$55.08_{1.63}$	$58.61_{1.20}$	49.11
	Hidd.C	$\mathbf{79.39}_{3.49}$	$\mathbf{67.23}_{1.13}$	$\mathbf{69.03}_{9.36}$	$\mathbf{33.12}_{2.59}$	$\mathbf{74.89}_{1.75}$	$\mathbf{52.50}_{2.17}$	$\mathbf{61.88}_{1.74}$	$\mathbf{46.63}_{5.67}$	$\mathbf{76.83}_{9.16}$	$73.14_{11.06}$	**63.46**
k=8	None	$9.18_{0.01}$	$53.56_{0.99}$	$55.06_{0.70}$	$26.48_{0.16}$	$35.11_{0.00}$	$35.06_{0.43}$	$37.56_{0.00}$	$37.93_{0.81}$	$36.12_{0.27}$	$88.65_{2.13}$	41.47
	Con.C	$11.00_{1.59}$	$52.04_{0.51}$	$54.55_{0.11}$	$27.01_{0.23}$	$35.11_{0.00}$	$34.60_{0.17}$	$37.56_{0.00}$	$38.81_{0.29}$	$38.68_{0.74}$	$\mathbf{89.87}_{0.41}$	41.92
	Bat.C	$18.25_{0.02}$	$52.17_{0.44}$	$54.91_{0.11}$	$25.89_{0.60}$	$\mathbf{66.77}_{0.63}$	$35.08_{0.61}$	$57.00_{0.26}$	$36.73_{0.86}$	$35.85_{0.25}$	$90.17_{1.13}$	47.28
	Dom.C	$9.39_{0.30}$	$53.02_{1.61}$	$55.18_{0.52}$	$\mathbf{32.92}_{1.16}$	$35.11_{0.00}$	$33.10_{2.08}$	$37.56_{0.00}$	$37.79_{0.23}$	$41.86_{1.38}$	$\mathbf{92.08}_{0.41}$	42.80
	KNN	$\mathbf{63.23}_{1.12}$	$\mathbf{60.79}_{0.34}$	$\mathbf{61.09}_{4.12}$	$30.84_{0.06}$	$63.67_{1.20}$	$\mathbf{37.53}_{0.81}$	$55.26_{5.26}$	$\mathbf{42.30}_{4.48}$	$\mathbf{80.93}_{0.92}$	$78.58_{0.41}$	**57.42**
	Cent.C	$57.73_{6.27}$	$57.02_{0.78}$	$54.73_{5.85}$	$26.35_{2.44}$	$62.27_{5.14}$	$35.94_{1.68}$	$\mathbf{59.23}_{6.62}$	$38.25_{6.66}$	$58.83_{10.95}$	$74.64_{0.80}$	52.50
	Hidd.C	$\mathbf{83.00}_{0.52}$	$\mathbf{68.71}_{0.42}$	$\mathbf{77.09}_{0.85}$	$\mathbf{31.59}_{8.39}$	$\mathbf{78.91}_{0.40}$	$\mathbf{60.97}_{0.90}$	$\mathbf{64.20}_{1.65}$	$\mathbf{50.52}_{0.85}$	$\mathbf{91.98}_{0.17}$	$84.78_{3.82}$	**69.17**