Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 264 (2025) 94–103

Procedia Computer Science

www.elsevier.com/locate/procedia

International Neural Network Society Workshop on Deep Learning Innovations and Applications

# N2C2: Nearest Neighbor Enhanced Confidence Calibration for Cross-Lingual In-Context Learning

Jie He[a], Simon Yu[a], Víctor Gutiérrez Basulto[b], Jeff Z. Pan[a]

[a]School of Informatics, University of Edinburgh, Edinburgh, UK
[b]School of Computer Science and Informatics, Cardiff University, Cardiff, UK

**Abstract**

Recent advancements of in-context learning (ICL) show language models can significantly improve their performance when demonstrations are provided. However, little attention has been paid to model calibration and prediction confidence of ICL in cross-lingual scenarios. To bridge this gap, we conduct a thorough analysis of ICL for cross-lingual sentiment classification. Our findings suggest that ICL performs poorly in cross-lingual scenarios, exhibiting low accuracy and presenting high calibration errors. In response, we propose a novel approach, **N2C2**, which employs a *k*-nearest neighbors augmented classifier for prediction confidence calibration. N2C2 narrows the prediction gap by leveraging a datastore of cached few-shot instances. Specifically, N2C2 integrates the predictions from the datastore and incorporates confidence-aware distribution, semantically consistent retrieval representation, and adaptive neighbor combination modules to effectively utilize the limited number of supporting instances. Evaluation on two multilingual sentiment classification datasets demonstrates that N2C2 outperforms traditional ICL. It surpasses fine tuning, prompt tuning and recent state-of-the-art methods in terms of accuracy and calibration errors.



*Keywords:* In-context learning, Multilingual Calibration, K-Nearest neighbors.

## 1. Introduction

In-context learning (ICL) has significantly enhanced the performance of pre-trained language models (PLMs) on few-shot tasks [3, 9, 23]. For instance, in cross-lingual learning, high-resource languages (e.g. English) can be explored to address data-scarce challenges in low-resource languages [17, 34, 31, 44].

Despite the progress made in ICL and its successful applications in cross-lingual scenarios, there is a noticeable gap in understanding the reliability of ICL methods in such tasks, i.e. how reliable their confidence predictions are.

* Corresponding author. Jie He
*E-mail address:* j.he@ed.ac.uk


10.1016/j.procs.2025.07.121

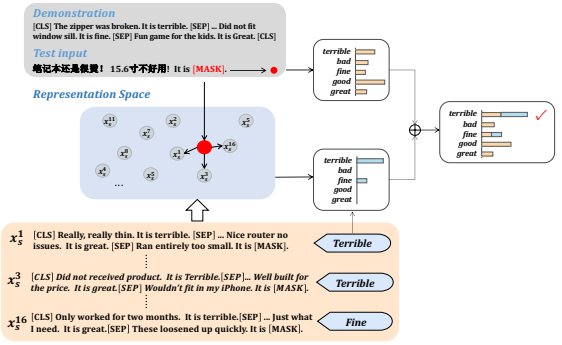


Fig. 1. Illustration of cross-lingual nearest neighbor inference with $k = 3$, which makes a prediction from 5 candidate words.

Specifically, there is a gap in assessing the consistency between model confidence and accuracy across different multilingual models [19, 41, 52]. Previous works in multilingual ICL often construct prompt-contexts using randomly selected input-label pairs [20, 37, 45]. While some efforts have been made to improve context selection in cross-lingual ICL [34], the focus has predominantly been on performance enhancement rather than reliability. One exception is the study conducted by [1], which investigates classical calibration techniques, such as temperature scaling, to mitigate calibration errors in multilingual classification tasks. However, that work concentrates on the calibration of fine-tuned multilingual models, yet it does not explore how this impacts the efficacy of cross-lingual ICL.

To bridge this gap, we start by assessing the performance of ICL on the cross-lingual sentiment classification (CLS) dataset [29]. In this setting, the context is provided in English while accuracy and calibration errors are evaluated on unseen languages. To measure the calibration error, we leverage the *expected calibration error* (ECE) [24], which measures the difference between confidence and accuracy, thus evaluating the model's reliability in cross-lingual ICL scenarios. Our study on CLS reveals subpar performance achieved by conventional ICL in both accuracy and calibration errors.

To tackle this challenge, we propose N2C2, a method that enhances the accuracy of conventional ICL while reducing the expected calibration error. N2C2 identifies examples in the source language, which support predictions in the target language (Fig. 1) based on three criteria: (a) accurate retrieval, (b) robustness of retrieved examples with limited training data, and (c) preference for retrieved examples with higher confidence.

To achieve criterion (a), N2C2 stores masked representations of each example in the source language, which are then transformed into retrieval-specific representations of lower dimensionality. For criterion (b), we develop a dynamic weighting mechanism in a trained neural network to adjust the importance of rertrieved examples, rather than relying on a fixed top-$K$ retrieval approach. To satisfy criterion (c), we integrate confidence into the probabilities derived from retrieval, rather than solely relying on distances obtained from retrieval. We evaluate N2C2 on the MARC and CLS datasets, with experimental results showcasing its superiority over existing baselines across various settings.

## 2. Related Work

**Calibration** Recent efforts have been made on the calibration of pre-trained language models [10, 6, 15, 9, 29, 2, 4, 46]. Particularly relevant are investigations by [1, 13], which explore the performance of various existing post-training calibration methods in cross-lingual classification and structure prediction tasks. Additionally, the study on contextual calibration by [42] calibrates ICL predictions through bias probing and conditional prediction reversal. However, all these works do not evaluate multilingual models' calibration under the ICL setting.

**Multilingual Prompt Learning** Reference [37] demonstrate the multilingual capabilities of language models trained on English data by using a few English examples as context and evaluating their performance on non-English data. Recent studies optimize prompts for cross-lingual ICL with multilingual PLMs [12, 41, 26, 34, 47, 51]. However, their focus is on retrieving useful demonstrations from the source language and concatenate them with target examples to enhance cross-lingual ICL performance. In contrast, our approach retrieves and utilizes labeled training examples to aid predictions on target samples.

**Retrieval in In-Context Learning** Previous studies show the benefits of selecting demonstration examples closely resembling the test input, particularly when ample training data is available [7, 21, 22, 48, 50]. Reference [21] retrieve the nearest training examples to the test input, employing unsupervised and supervised methods. Reference [23] and [31] use nearest neighbor search to incorporate additional data for zero-shot inference by retrieving sentences closely related to the test input. Closer to our approach, Reference [27] and [39] enhance monolingual classification using $k$NN retrieval, but they do not address calibration in retrieval-augmented multilingual models. In contrast, our proposed retrieval-augmented approach is specifically designed for cross-lingual in-context learning.

## 3. Background

**Task Formulation** Our key interest is cross-lingual ICL. Let $s$ be a source language, we use $X_s$ and $Y_s$ to respectively denote the sets of input examples and their corresponding labels in $s$. We consider a monolingual labeled dataset $D_s = \{(x_i^s, y_i^s)\}_{i=1}^m$ with $m$ sampled examples, where $x_i^s \in X_s$ and $y_i^s \in Y_s$. Let $t$ be a target language, with $X_t$ as above. We consider the set of sampled examples $D_t = \{x_j^t\}_{j=1}^n$, with $n$ samples from $X_t$. For cross-lingual ICL, we randomly choose **1**-shot per-class input-label pairs from $D_s$ as a prompt-context $C_s$: $C_s = \pi(x_1^s, y_1^s)$ , where $\pi$ is the prompt template. The cross-lingual context $C_s$ is concatenated with the prompted input to form the input $I_j$ for the multilingual PLM:

$$I_j = C_s \oplus \pi(x_j^t, \texttt{[MASK]}) \tag{1}$$

$$\pi(x_j^t, \texttt{[MASK]}) = x_j^t. \texttt{ It is [MASK].} \tag{2}$$

where $\oplus$ is the concatenation operator. The multilingual PLM is responsible of predicting masked tokens in the input $I_j$ and providing probability estimates $p$ for all possible candidate words. For a candidate label $y$ over the label space $\mathcal{Y}$, we determine the predicted class $\hat{y}$ by selecting the verbalizer $v(y)$ with the highest probability.

**Ensemble-based Cross-lingual ICL** We mainly use the XLM-RoBERTa (XLM-R) model, which is limited by the input length, so we uniformly apply an ICL ensemble method across different shot settings [13]. This approach involves partitioning $D_s$ into $m > 0$ demonstration sets $D_s = D_1 \cup D_2 \cup ... \cup D_m$ to create different prompt contexts and combining the predictions from these $m$ prompts. Specifically, for calculating the prediction of the target example $x_j^t$, we compute $\frac{1}{m} \sum_{i=1}^m P(y|x_i^s, y_i^s, x_j^t)$.

**Calibration for Cross-lingual ICL** Calibration refers to the alignment between a model's assigned probability (*confidence*) for a prediction and the true measure of its correctness (*accuracy*) [25]. In other words, given an input $x$, the ground truth $y$ and a prediction $\hat{y}$, the *perfectly* calibrated confidence $\text{conf}(x, \hat{y})$ will satisfy: $\forall p \in [0, 1], P(\hat{y} = y \mid \text{conf}(x, \hat{y}) = p) = p$. The *expected calibration error (ECE)* is a widely used metric to asses miscalibration that quantifies the difference between the expected confidence and accuracy [24].

## 4. Our Method: N2C2

**Overview.** Our method consists of **four steps**, cf. Figure 2. **First**, the masked representations generated by multilingual pre-trained language models (MPLMs) are used to represent sentences after passing them through an MLP layer for semantically consistent retrieval representation (§4.2). **Second**, we retrieve a training set by finding the $k$-nearest examples (§4.1). **Third**, we assign higher weights to the examples with higher prediction confidence, to assist in the test input prediction (§4.3). **Finally**, to enhance the robustness of our method, instead of relying solely on a fixed number of top-$K$ predictions, we divide $K$ into $\{K_1, K_2, ..., K_l\}$ and train a lightweight network to merge the results from these different $K_i$ (§4.4). For training these lightweight networks, we split the training set into two equal parts: one for retrieval and another one for updating the modules.

We explain N2C2's steps through an example, following Fig. 2. **Task**: Binary sentiment classification, labels 0, 1. **Input:** A target language test example $x^t$ and source examples $x_i^s$, $i \in [1, 6]$ with corresponding labels 0, 1, 1, 0, 1, 0. In Step (1), the MASK representation of the test example is transformed nonlinearly to obtain the representation **h**. In Step

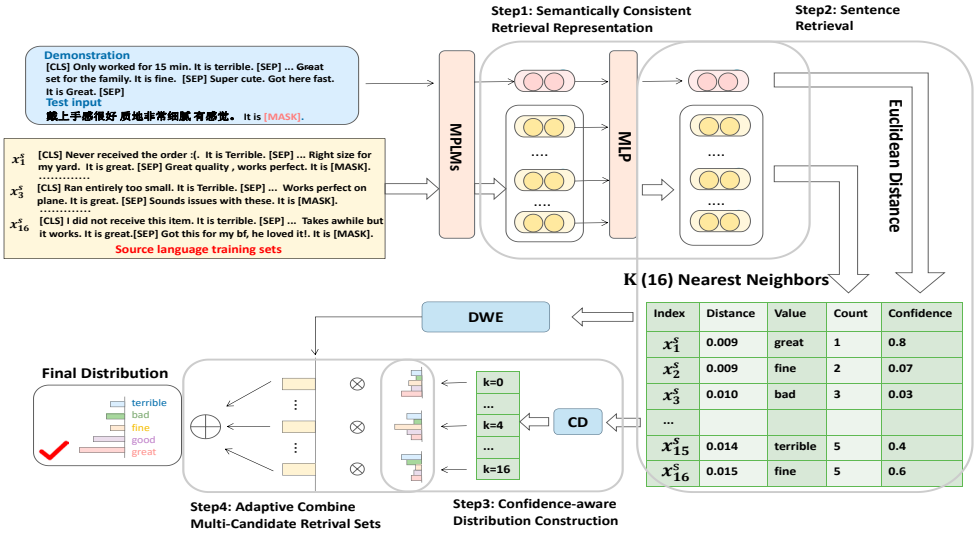


Fig. 2. Diagram of N2C2 with $k = 16$. N2C2 first reconstructs $\mathrm{h}_{[\mathrm{mask}]}$(§4.2) for the test example in the target language, and selects neighbors (§4.1) for it. It then consider confidence to generate multiple distributions (§4.3). These distributions are summed up together to form the final predicted distribution (§4.4).

(2), we calculate the similarities $s_i$, $i \in [1, 6]$ between $\mathbf{h}$ and the representations $x_i^s$. We will only use the top 4 retrieved examples, so we only consider $s_1, s_2, s_3, s_4$ (which are the top 4 examples here) in the next two steps. In Step (3), we consider the model's confidence in the predictions for these four retrieved examples together with their distance to refine the generated retrieval distribution: $distri_4 = [p_0, p_1]$, where $distri_4$ denotes the nearest top 4 retrieved examples and $p_i$, $i \in \{0, 1\}$, a probability class. In Step (4), to enhance flexibility and robustness, we consider the top 4 examples (note that 4 is a hyperparemeter) and additionally the top 2. This yields answer distributions $dis_2$ and $dis_4$, respectively. We additionally train aweighting module to gauge the importance of different answer distributions. The final predicted answer distribution is obtained as $w_1 \times distri_2 + w_2 \times distri_4$.

### 4.1. Inference with kNN Retrieval

To do inference on a test example $x_j^t$, we use Equation (2) to obtain the vector representation $\boldsymbol{h}_j^t$, corresponding to its `[MASK]` position. Subsequently, we derive the $k$NN-based prediction distribution $\boldsymbol{p}_{k\mathrm{NN}}$ over the label space $\mathcal{Y}$ by considering the nearest neighbors:

$$p_{k\mathrm{NN}}(y|x_j^t) \propto \sum_{(h_i^s, y_i^s) \in N_t} \mathbb{1}_{y=y_i} \exp\left(\frac{-d(h_i^s, h_j^t)}{\tau}\right) \tag{3}$$

where $d$ denotes the Euclidean distance, $N_t$ denotes the set of $K$ nearest neighbors, and $\tau$ is a temperature parameter used to control the sharpness of the *softmax* function. The final prediction distribution for label $y$ is obtained by interpolating two distributions, with the interpolation factor $\lambda$ being tuned within the range of [0, 1]:

$$p(y|x_j^t) = (1 - \lambda) \cdot p_{k\mathrm{NN}}(y|x_j^t) + \lambda \cdot p_{\mathrm{MPLM}}\left(y|x_j^t\right)$$

Additionally, following the concept of ensemble ICL (§3), we aggregate the outcomes from various demonstrations for each test instance during the inference stage.

### 4.2. Semantically Consistent Retrieval Representation

We esttimate the representation $\boldsymbol{h_j^t}$ of example $x_j^t$ using Eq. (2). During the inference phase, for a test instance, we use its representations to retrieve the top $K$ instances from the source training language, along with their corresponding labels. However, to deal with a situation in which retrieved instances with similar representations have different labels, we incorporate a straightforward linear layer that helps to distinguish such instance representations by leveraging the supervision provided by all training instances: $h_j^{t'} = \mathbf{W}^T h_j^t + b$, where $\mathbf{W} \in \mathbb{R}^{H \times Z}$ and $b \in \mathbb{R}^Z$ are trainable parameters, with $Z$ representing the new dimension of the representation space. The purpose of this linear module is to ensure, by maximizing the $k$NN retrieval probability defined in Eq. (3), that representations belonging to the same class are semantically similar. To achieve this, we optimize the linear layer by minimizing the cross-entropy loss associated with Eq. (3).

### 4.3. Confidence-Aware Distribution Construction

After computing the $k$NN distribution with $\boldsymbol{h_j^{t'}}$ using Eq. (3), we observe that the weights assigned to the retrieved examples are solely based on their distance to the query. However, this approach is suboptimal as it does not consider the confidence of the model. To address this, we introduce the *confidence-aware distribution (CD)* module, which assesses the importance of each retrieved pair $(x_i^s, y_i^s)$. This assessment is then used to refine the $k$NN distribution, ensuring that the probabilities are adjusted accordingly. Specifically, the $k$NN distribution for example $x_j^t$ is constructed through the following process:

$$p_{k\text{NN}}(y|x_j^t) \propto \sum_{(h_i,y_i)\in N_t} \mathbb{1}_{y=y_i} \exp\left(\frac{-d(h_i, h_j^{t'})}{\tau \times T} + C\right) \tag{4}$$

$$T = \mathbf{W}_1(\tanh(\mathbf{W}_2[d_1, \ldots, d_M; o_1, \ldots, o_M])) \tag{5}$$

$$C = \mathbf{W}_3(\tanh(\mathbf{W}_4[p(y_i^s|x_j^t); p(y_i^s|x_i^s)])) \tag{6}$$

In the calculation of $C$ we consider two types of information: 1) $p(y_i^s|x_j^t)$, which denotes the predicted probability of $y_i^s$ by the multilingual PLM, given the mask representation $\boldsymbol{h_j^t}$, and 2) $p(y_i^s|x_i^s)$, which denotes the predicted probability of $y_i^s$ for the retrieved sample $x_i^s$. The variable $\boldsymbol{d_i}$ represents the L2 distance between the query $\boldsymbol{h_j^t}$ and the retrieved sample $\boldsymbol{h_i^s}$, while $\boldsymbol{o_i}$ denotes the number of unique values among the top $i$ neighbors. The $\mathbf{W}$s are the the parameter matrices.

### 4.4. Adaptively Combining Multi-Candidate Retrieval Sets

Using a fixed value of $K$ in Eq. (3) can be problematic, especially when there are insufficient relevant items in the training sets. To enhance the robustness of N2C2, we adopt a technique proposed by [43, 49]. This approach involves considering a range of $K$ values that are smaller than a predefined upper bound, $K_{max}$. Additionally, we introduce a lightweight network to assess the importance of using different selections, thereby minimizing the inclusion of irrelevant neighbors. In practice, we simplify the choice of $K$ by opting for multiples of 4 (denoted as $R_s$), which includes $K = 0$ (corresponding to solely utilizing multilingual PLMs). The lightweight network evaluates the importance of various $k$NN retrieval outcomes by using the retrieved neighbors as inputs.

More specifically, for a given test instance with a demonstration $x_j^t$, we begin by retrieving a maximum of $K_{max}$ neighbors from the source language $s$. Then, we compute the distances between these neighbors and the current representation, as well as the count of distinct values in the top $i$ neighbors denoted as $o_i$. The calculated distances $d = (d_1, ..., d_{K_{max}})$ and counts $c = (c_1, ..., c_{K_{max}})$ serve as inputs for determining the optimal value of $K$. The rationale behind considering the distances of each neighbor lies in the direct evidence they provide for assessing their importance.

Furthermore, we incorporate the label counts $\boldsymbol{o_i}$, following Eq. (5). The distribution weight estimation network, denoted as $\mathbf{DWE}(\cdot)$, comprises two feed-forward layers with a non-linear function in between. Specifically, we set

the hidden size to 32. The probability of selecting a particular value of $K$ is calculated using the following formula: $p(K|x_j^t) = softmax(f_{\text{DWE}}([d, c]))$. Instead of relying solely on the hyperparameter $\lambda$ as defined in Equation 4.1, we employ the importance estimation network to combine the outputs of PLMs and various $k$NN predictions as the final prediction:

$$p(y|x_j^t) = \sum_{M \in R_s} p(M|x_j^t) \cdot p_{k\text{NN}_\text{M}}(y|x_j^t) \tag{7}$$

## 5. Experiments

**Datasets** We evaluated N2C2 and baselines on two datasets: Multilingual Amazon Reviews Corpus (MARC) [16] and Cross-language sentiment classification (CLS) [29].

Table 1. Main results for the baselines and our method, reported as **Accuracy** ↑ (%) / **ECE** ↓ (%). $b$ is the number of training samples per class (i.e. $b$-shot). "Avg." is the average result for all languages. We report the mean results across 20 runs with random restarts. The subscript represents the corresponding standard deviation.

| $b$ | Lang | ICL | ICL + CC | FT | PT | X-InSTA | X-InSTA* | N2C2 |
|---|---|---|---|---|---|---|---|---|
| | | | | **MARC** | | | | |
| **2** | De | $28.06_{\pm 3.2}$ / $28.00_{\pm 8.0}$ | $27.84_{\pm 3.1}$ / $19.00_{\pm 5.0}$ | $21.28_{\pm 0.4}$ / $53.92_{\pm 18.4}$ | $26.53_{\pm 1.5}$ / $55.70_{\pm 2.4}$ | 20.36 / 71.65 | 27.84 / 80.54 | **29.09**$_{\pm 2.0}$ / **14.97**$_{\pm 4.4}$ |
| | En | $32.42_{\pm 4.0}$ / $23.00_{\pm 7.0}$ | $31.16_{\pm 3.8}$ / $22.00_{\pm 5.0}$ | $20.96_{\pm 0.3}$ / $55.53_{\pm 20.9}$ | $31.70_{\pm 1.9}$ / $56.72_{\pm 2.0}$ | 22.12 / 57.89 | **43.66** / 69.51 | $32.56_{\pm 2.7}$ / **11.59**$_{\pm 3.7}$ |
| | Zh | $24.90_{\pm 2.6}$ / $32.00_{\pm 9.0}$ | $25.57_{\pm 2.2}$ / $22.00_{\pm 3.0}$ | $22.37_{\pm 1.0}$ / $54.00_{\pm 19.9}$ | $22.06_{\pm 0.9}$ / $59.60_{\pm 2.7}$ | 19.24 / 54.38 | **36.14** / 39.79 | $26.79_{\pm 1.8}$ / **15.71**$_{\pm 5.0}$ |
| | Avg. | $27.16_{\pm 2.7}$ / $30.00_{\pm 4.4}$ | $27.45_{\pm 1.8}$ / $20.50_{\pm 1.3}$ | $21.37_{\pm 0.7}$ / $54.57_{\pm 2.6}$ | $26.48_{\pm 2.8}$ / $57.03_{\pm 1.3}$ | 20.51 / 60.82 | **33.90** / 52.25 | $28.74_{\pm 1.9}$ / **15.57**$_{\pm 2.7}$ |
| **4** | De | $26.58_{\pm 3.3}$ / $32.00_{\pm 9.0}$ | $28.64_{\pm 2.5}$ / $22.00_{\pm 6.0}$ | $27.98_{\pm 1.9}$ / $53.03_{\pm 15.7}$ | $26.57_{\pm 1.5}$ / $62.82_{\pm 6.1}$ | 19.58 / 73.72 | 27.12 / 78.07 | **34.13**$_{\pm 3.0}$ / **7.94**$_{\pm 2.8}$ |
| | En | $31.58_{\pm 4.8}$ / $25.00_{\pm 8.0}$ | $32.04_{\pm 3.4}$ / $24.00_{\pm 6.0}$ | $26.15_{\pm 1.4}$ / $51.00_{\pm 23.7}$ | $27.80_{\pm 1.7}$ / $62.56_{\pm 6.0}$ | 22.24 / 56.39 | **45.24** / 73.73 | $36.65_{\pm 3.4}$ / **5.73**$_{\pm 2.4}$ |
| | Zh | $23.33_{\pm 2.3}$ / $36.00_{\pm 10.0}$ | $27.97_{\pm 2.0}$ / $25.00_{\pm 6.0}$ | $22.45_{\pm 1.3}$ / $48.96_{\pm 24.3}$ | $23.67_{\pm 2.3}$ / $65.64_{\pm 7.3}$ | 20.48 / 50.64 | **36.68** / 39.34 | $32.50_{\pm 2.9}$ / **7.80**$_{\pm 3.1}$ |
| | Avg. | $25.59_{\pm 3.0}$ / $33.67_{\pm 4.5}$ | $29.32_{\pm 1.3}$ / $22.67_{\pm 1.6}$ | $24.39_{\pm 2.6}$ / $53.56_{\pm 4.4}$ | $25.17_{\pm 1.6}$ / $64.32_{\pm 1.2}$ | 20.62 / 60.33 | **34.07** / 52.51 | $33.31_{\pm 1.7}$ / **8.09**$_{\pm 1.6}$ |
| **8** | De | $27.38_{\pm 4.0}$ / $29.00_{\pm 10.0}$ | $28.91_{\pm 2.4}$ / $23.00_{\pm 8.0}$ | $25.70_{\pm 0.7}$ / $53.83_{\pm 17.2}$ | $31.04_{\pm 1.9}$ / $62.19_{\pm 5.5}$ | 19.68 / 71.66 | 26.48 / 76.32 | **35.08**$_{\pm 1.8}$ / **15.07**$_{\pm 3.2}$ |
| | En | $32.37_{\pm 4.4}$ / $22.00_{\pm 8.0}$ | $32.81_{\pm 2.8}$ / $24.00_{\pm 8.0}$ | $30.62_{\pm 1.2}$ / $57.99_{\pm 7.8}$ | $33.10_{\pm 1.1}$ / $61.34_{\pm 4.4}$ | 22.20 / 57.01 | **45.14** / 77.55 | $39.18_{\pm 1.3}$ / **9.87**$_{\pm 2.2}$ |
| | Zh | $23.82_{\pm 2.8}$ / $34.00_{\pm 11.0}$ | $27.97_{\pm 1.2}$ / $27.00_{\pm 8.0}$ | $26.53_{\pm 0.8}$ / $59.77_{\pm 6.7}$ | $28.48_{\pm 1.4}$ / $64.80_{\pm 5.1}$ | 22.30 / 52.66 | **37.10** / 39.07 | $33.05_{\pm 1.7}$ / **13.69**$_{\pm 2.6}$ |
| | Avg. | $26.20_{\pm 3.1}$ / $31.00_{\pm 4.8}$ | $29.61_{\pm 1.5}$ / $23.33_{\pm 2.0}$ | $28.59_{\pm 1.9}$ / $55.20_{\pm 3.3}$ | $30.18_{\pm 1.7}$ / $63.25_{\pm 1.3}$ | 20.91 / 60.33 | 33.82 / 52.54 | **34.36**$_{\pm 2.4}$ / **15.09**$_{\pm 2.9}$ |
| **16** | De | $27.36_{\pm 4.0}$ / $27.00_{\pm 10.0}$ | $30.61_{\pm 3.3}$ / $21.00_{\pm 7.0}$ | $36.92_{\pm 2.6}$ / $48.69_{\pm 17.4}$ | $36.46_{\pm 2.0}$ / $56.39_{\pm 4.8}$ | 19.68 / 64.07 | 25.62 / 76.07 | **37.75**$_{\pm 2.3}$ / **10.94**$_{\pm 5.9}$ |
| | En | $32.20_{\pm 4.3}$ / $21.00_{\pm 7.0}$ | $32.71_{\pm 2.6}$ / $17.00_{\pm 5.0}$ | $28.42_{\pm 1.0}$ / $50.99_{\pm 25.4}$ | $37.00_{\pm 1.5}$ / $56.78_{\pm 3.9}$ | 21.34 / 52.38 | **44.00** / 76.48 | $42.85_{\pm 1.4}$ / **5.69**$_{\pm 2.7}$ |
| | Zh | $23.64_{\pm 2.5}$ / $31.00_{\pm 10.0}$ | $28.28_{\pm 1.8}$ / $25.00_{\pm 7.0}$ | $32.53_{\pm 1.8}$ / $57.10_{\pm 6.5}$ | $31.82_{\pm 2.3}$ / $59.84_{\pm 5.7}$ | 23.16 / 55.45 | **36.96** / 38.88 | $35.10_{\pm 2.1}$ / **9.49**$_{\pm 4.5}$ |
| | Avg. | $26.16_{\pm 3.1}$ / $28.67_{\pm 4.1}$ | $30.23_{\pm 1.4}$ / $20.17_{\pm 2.6}$ | $31.78_{\pm 2.7}$ / $53.05_{\pm 3.2}$ | $33.79_{\pm 2.2}$ / $57.94_{\pm 1.3}$ | 20.91 / 58.94 | 33.05 / 51.66 | **37.07**$_{\pm 3.2}$ / **10.89**$_{\pm 3.5}$ |
| **32** | De | $27.55_{\pm 4.0}$ / $19.00_{\pm 10.0}$ | $22.53_{\pm 3.2}$ / $21.00_{\pm 6.0}$ | $38.78_{\pm 1.5}$ / $55.72_{\pm 2.7}$ | $39.16_{\pm 2.4}$ / $54.94_{\pm 3.2}$ | 19.92 / 60.24 | 24.54 / 75.80 | **41.59**$_{\pm 1.8}$ / **4.35**$_{\pm 2.2}$ |
| | En | $32.15_{\pm 4.5}$ / $19.00_{\pm 8.0}$ | $22.57_{\pm 2.4}$ / $18.00_{\pm 4.0}$ | $39.31_{\pm 1.0}$ / $54.45_{\pm 4.3}$ | $40.51_{\pm 1.9}$ / $54.56_{\pm 2.5}$ | 19.92 / 60.24 | 43.42 / 75.22 | **44.79**$_{\pm 1.2}$ / **2.72**$_{\pm 1.3}$ |
| | Zh | $23.81_{\pm 2.6}$ / $21.00_{\pm 9.0}$ | $21.31_{\pm 1.9}$ / $25.00_{\pm 7.0}$ | $36.99_{\pm 2.1}$ / $53.03_{\pm 6.4}$ | $35.59_{\pm 2.6}$ / $58.23_{\pm 3.7}$ | 18.32 / 69.26 | 37.06 / 39.02 | **39.87**$_{\pm 2.1}$ / **4.12**$_{\pm 2.6}$ |
| | Avg. | $26.20_{\pm 3.0}$ / $19.00_{\pm 1.5}$ | $22.38_{\pm 0.6}$ / $20.17_{\pm 2.4}$ | $36.90_{\pm 1.7}$ / $54.08_{\pm 1.8}$ | $37.50_{\pm 1.8}$ / $55.84_{\pm 1.3}$ | 22.04 / 49.73 | 32.57 / 51.16 | **40.72**$_{\pm 2.3}$ / **4.57**$_{\pm 1.0}$ |
| | | | | **CLS** | | | | |
| **2** | En | $29.55_{\pm 8.2}$ / $27.00_{\pm 10.0}$ | $30.23_{\pm 7.4}$ / $37.00_{\pm 11.0}$ | $25.03_{\pm 6.6}$ / $40.57_{\pm 21.7}$ | $33.18_{\pm 8.2}$ / $62.13_{\pm 11.7}$ | 15.40 / 49.19 | **52.15** / 73.01 | $30.46_{\pm 5.3}$ / **18.50**$_{\pm 5.9}$ |
| | Fr | $18.41_{\pm 3.6}$ / $40.00_{\pm 7.0}$ | $22.49_{\pm 5.2}$ / $43.00_{\pm 7.0}$ | $27.53_{\pm 5.2}$ / $40.99_{\pm 24.4}$ | $31.85_{\pm 9.2}$ / $62.41_{\pm 14.2}$ | **34.20** / 50.61 | 33.10 / 44.36 | $20.51_{\pm 3.1}$ / **29.43**$_{\pm 6.2}$ |
| | Avg. | $21.72_{\pm 4.6}$ / $36.25_{\pm 6.1}$ | $24.20_{\pm 3.5}$ / $42.75_{\pm 3.9}$ | $24.85_{\pm 2.1}$ / $42.15_{\pm 1.8}$ | $31.83_{\pm 1.2}$ / $62.49_{\pm 1.3}$ | 24.72 / 54.64 | **38.12** / 53.82 | $22.96_{\pm 4.3}$ / **26.66**$_{\pm 4.7}$ |
| **4** | En | $29.60_{\pm 7.6}$ / $22.00_{\pm 9.0}$ | $27.10_{\pm 4.0}$ / $38.00_{\pm 13.0}$ | $40.85_{\pm 2.3}$ / $40.72_{\pm 12.0}$ | $37.59_{\pm 4.8}$ / $56.13_{\pm 6.5}$ | 15.25 / 47.5 | **54.05** / 75.57 | $35.34_{\pm 4.2}$ / **15.88**$_{\pm 5.8}$ |
| | Fr | $21.78_{\pm 3.0}$ / $26.00_{\pm 6.0}$ | $24.02_{\pm 3.4}$ / $40.00_{\pm 13.0}$ | $41.18_{\pm 2.8}$ / $46.07_{\pm 7.5}$ | **39.44**$_{\pm 4.7}$ / $52.89_{\pm 6.5}$ | 34.90 / 53.75 | 34.85 / 42.69 | $27.29_{\pm 2.0}$ / **22.55**$_{\pm 2.6}$ |
| | Avg. | $23.50_{\pm 3.6}$ / $24.75_{\pm 3.0}$ | $24.50_{\pm 1.6}$ / $40.50_{\pm 1.8}$ | $39.24_{\pm 1.9}$ / $44.82_{\pm 2.5}$ | $38.52_{\pm 1.0}$ / $53.98_{\pm 1.3}$ | 24.91 / 56.98 | **39.11** / 53.56 | $28.68_{\pm 4.0}$ / **20.84**$_{\pm 3.0}$ |
| **8** | En | $27.57_{\pm 7.1}$ / $17.00_{\pm 8.0}$ | $27.23_{\pm 5.1}$ / $39.00_{\pm 18.0}$ | $37.24_{\pm 2.9}$ / $38.95_{\pm 9.5}$ | $38.69_{\pm 2.3}$ / $48.51_{\pm 1.4}$ | 15.15 / 46.36 | **52.60** / 76.30 | $42.85_{\pm 4.7}$ / **9.30**$_{\pm 4.1}$ |
| | Fr | $18.53_{\pm 3.3}$ / $26.00_{\pm 5.0}$ | $24.20_{\pm 6.0}$ / $44.00_{\pm 18.0}$ | $34.69_{\pm 2.0}$ / $44.83_{\pm 19.4}$ | $33.09_{\pm 2.7}$ / $49.12_{\pm 2.5}$ | 35.25 / 55.33 | **35.90** / 42.65 | $35.58_{\pm 3.3}$ / **13.22**$_{\pm 4.3}$ |
| | Avg. | $21.54_{\pm 3.7}$ / $23.00_{\pm 4.3}$ | $25.56_{\pm 1.1}$ / $42.00_{\pm 2.1}$ | $31.57_{\pm 4.6}$ / $41.94_{\pm 2.6}$ | $33.04_{\pm 1.6}$ / $49.76_{\pm 1.2}$ | 24.35 / 57.50 | **38.75** / 53.02 | $36.78_{\pm 3.7}$ / **12.46**$_{\pm 1.9}$ |
| **16** | En | $27.32_{\pm 7.2}$ / $31.00_{\pm 10.0}$ | $25.10_{\pm 4.3}$ / $39.00_{\pm 10.0}$ | $42.03_{\pm 2.0}$ / $50.20_{\pm 4.6}$ | $44.83_{\pm 4.7}$ / $51.44_{\pm 5.0}$ | 15.15 / 46.41 | **52.55** / 74.96 | $47.64_{\pm 2.5}$ / **6.56**$_{\pm 2.6}$ |
| | Fr | $18.67_{\pm 3.2}$ / $40.00_{\pm 7.0}$ | $21.80_{\pm 5.0}$ / $43.00_{\pm 11.0}$ | $39.31_{\pm 10.1}$ / $49.11_{\pm 22.9}$ | $38.48_{\pm 4.4}$ / $57.19_{\pm 5.1}$ | 35.60 / 55.20 | 34.90 / 42.62 | **41.64**$_{\pm 5.0}$ / **8.00**$_{\pm 4.6}$ |
| | Avg. | $21.37_{\pm 3.5}$ / $37.25_{\pm 4.6}$ | $22.65_{\pm 1.5}$ / $41.75_{\pm 1.6}$ | $36.00_{\pm 5.4}$ / $37.48_{\pm 17.6}$ | $40.85_{\pm 2.4}$ / $54.06_{\pm 2.3}$ | 24.92 / 57.52 | 38.56 / 52.60 | **43.99**$_{\pm 3.0}$ / **7.61**$_{\pm 0.6}$ |
| **32** | En | $26.69_{\pm 6.9}$ / $28.00_{\pm 9.0}$ | $22.53_{\pm 4.3}$ / $31.00_{\pm 6.0}$ | $51.27_{\pm 5.5}$ / $36.07_{\pm 11.2}$ | $48.98_{\pm 2.6}$ / $46.62_{\pm 3.3}$ | 15.15 / 45.12 | 51.45 / 74.91 | **55.63**$_{\pm 2.0}$ / **4.50**$_{\pm 1.3}$ |
| | Fr | $18.67_{\pm 3.3}$ / $37.00_{\pm 7.0}$ | $20.24_{\pm 3.9}$ / $32.00_{\pm 7.0}$ | $27.04_{\pm 14.6}$ / $14.07_{\pm 2.6}$ | $40.54_{\pm 3.3}$ / $51.42_{\pm 4.1}$ | 35.65 / 55.61 | 35.30 / 40.61 | **43.73**$_{\pm 3.7}$ / **4.83**$_{\pm 3.4}$ |
| | Avg. | $21.25_{\pm 3.2}$ / $34.00_{\pm 4.3}$ | $20.92_{\pm 1.3}$ / $31.75_{\pm 0.8}$ | $39.95_{\pm 8.9}$ / $35.38_{\pm 13.1}$ | $44.42_{\pm 3.1}$ / $49.33_{\pm 1.8}$ | 24.48 / 57.32 | 38.29 / 51.86 | **47.27**$_{\pm 4.9}$ / **5.31**$_{\pm 0.7}$ |

**Baselines** We compared N2C2 with the following cross-lingual language models: (1) **In-Context Learning (ICL)** [3]: this method utilizes $b$ input-label pairs from the training data and employs in-context learning. (2) **Contextual Calibration (ICL+CC)** [42]: this approach addresses prediction bias in ICL by introducing a content-free input "N/A"; (3) **Fine Tuning (FT)**: this method uses a classifier head that takes the [CLS] token as input and fine-tunes over MPLMs with the classification head; (4) **Prompt Tuning (PT)** [30]: a prompt-based fine-tuning method that utilizes manual prompts and fine-tunes over MPLMs. (5) **Cross-lingual In-context Source-Target Alignment (X-InSTA)** [34]: it prepends the top-$K$ similar in-context samples which are retrieved from the training samples augmented with task alignment.

Table 2. **Ablation study results for N2C2.** Results are reported as **Accuracy** ↑ (%) / **ECE** ↓ (%). The p-value is calculated by two-tailed t-tests.

| Language | N2C2 | (1) W/o CD | (2) W/o retrieval representation shaping | (3) W/o DWE |
|---|---|---|---|---|
| | | **MARC** | | |
| De | $37.75_{\pm 2.3}$ / $10.94_{\pm 5.9}$ | $37.07_{\pm 2.4}$ / $12.35_{\pm 5.9}$ | $36.16_{\pm 1.2}$ / $4.82_{\pm 1.0}$ | $37.99_{\pm 2.5}$ / $16.18_{\pm 6.4}$ |
| En | $42.85_{\pm 1.4}$ / $5.69_{\pm 2.7}$ | $42.10_{\pm 1.6}$ / $6.90_{\pm 1.8}$ | $38.09_{\pm 1.9}$ / $4.91_{\pm 1.2}$ | $43.33_{\pm 1.4}$ / $11.06_{\pm 3.6}$ |
| Es | $34.50_{\pm 2.0}$ / $12.88_{\pm 5.4}$ | $33.76_{\pm 1.9}$ / $14.45_{\pm 5.0}$ | $33.95_{\pm 1.6}$ / $4.49_{\pm 1.6}$ | $34.90_{\pm 2.2}$ / $17.45_{\pm 6.1}$ |
| Fr | $33.51_{\pm 2.0}$ / $17.01_{\pm 6.4}$ | $32.92_{\pm 1.7}$ / $18.50_{\pm 6.3}$ | $33.17_{\pm 1.9}$ / $5.72_{\pm 1.9}$ | $34.06_{\pm 2.0}$ / $22.30_{\pm 7.1}$ |
| Ja | $38.72_{\pm 2.3}$ / $9.36_{\pm 4.8}$ | $37.98_{\pm 1.8}$ / $10.37_{\pm 3.6}$ | $34.85_{\pm 2.1}$ / $5.26_{\pm 2.2}$ | $39.37_{\pm 2.1}$ / $14.92_{\pm 6.0}$ |
| Zh | $35.10_{\pm 2.1}$ / $9.49_{\pm 4.5}$ | $34.39_{\pm 1.5}$ / $10.62_{\pm 3.5}$ | $33.45_{\pm 1.4}$ / $4.82_{\pm 1.4}$ | $35.74_{\pm 2.1}$ / $14.17_{\pm 5.3}$ |
| Avg. | $37.07_{\pm 3.2}$ / $10.89_{\pm 3.5}$ | $36.37_{\pm 3.4}$ / $12.20_{\pm 4.0}$ | $34.94_{\pm 1.9}$ / $5.00_{\pm 0.4}$ | $37.56_{\pm 3.5}$ / $16.01_{\pm 3.8}$ |
| *p-value* | - | 8.6e-01 / 1.0e-02 | 3.4e-02 / 1.1e-02 | 5.8e-04 / 6.1e-07 |
| | | **CLS** | | |
| De | $46.15_{\pm 6.9}$ / $7.84_{\pm 4.0}$ | $45.02_{\pm 7.2}$ / $8.40_{\pm 4.3}$ | $44.21_{\pm 6.6}$ / $7.19_{\pm 3.6}$ | $45.85_{\pm 7.0}$ / $11.79_{\pm 5.4}$ |
| En | $47.64_{\pm 2.5}$ / $6.56_{\pm 2.6}$ | $46.37_{\pm 2.3}$ / $7.88_{\pm 2.6}$ | $43.41_{\pm 4.1}$ / $5.19_{\pm 2.5}$ | $48.99_{\pm 2.4}$ / $11.52_{\pm 4.5}$ |
| Fr | $41.64_{\pm 5.0}$ / $8.00_{\pm 4.6}$ | $40.71_{\pm 5.4}$ / $9.09_{\pm 4.6}$ | $41.20_{\pm 5.2}$ / $5.88_{\pm 2.1}$ | $41.01_{\pm 4.8}$ / $9.79_{\pm 5.5}$ |
| Jp | $40.52_{\pm 3.2}$ / $8.05_{\pm 2.9}$ | $39.66_{\pm 3.1}$ / $8.82_{\pm 2.4}$ | $33.65_{\pm 2.3}$ / $6.25_{\pm 2.5}$ | $40.87_{\pm 3.0}$ / $14.66_{\pm 4.4}$ |
| Avg. | $43.99_{\pm 3.0}$ / $7.61_{\pm 0.6}$ | $42.94_{\pm 3.3}$ / $8.55_{\pm 0.5}$ | $40.62_{\pm 4.8}$ / $6.13_{\pm 0.8}$ | $44.18_{\pm 3.0}$ / $11.94_{\pm 2.0}$ |
| *p-value* | - | 1.7e-01 / 3.6e-02 | 9.6e-02 / 1.9e-02 | 6.9e-01 / 2.3e-02 |

To ensure fairness, we used XLM-R as the base model, but we also present results for the variant X-InSTA*, using XGLM-7.5B [20].

### *5.1. Main Results*

Table 1 offers an overview of the performance of N2C2 alongside the baselines. From Table 1, it is evident that N2C2 outperforms all baselines in terms of accuracy. N2C2 performs very well across all datasets, achieving an average 4.2%, 3.24% accuracy improvement over the strong baseline (PT) on the MARC and CLS datasets, respectively, with only 8, 16 and 32 shots. In addition, it is noteworthy that that the effectiveness of the X-InSTA model does not increase with the number of shots, whereas our method shows a continuous improvement as the number of shots increases, even surpassing X-InSTA* when $b = 32$.

Particularly remarkable is N2C2's performance in terms of ECE, with an average decrease of 10.53% and 16.47% when compared to the strongest baseline on the MARC and CLS datasets, respectively. Another important observation is that while the baseline methods show varying degrees of improvement in accuracy with the increase of training data, their ECE does not consistently decrease and in some cases even exhibits a slight increase. In contrast, our proposed method demonstrates a consistent decrease in ECE.

Only when compared to X-InSTA*, which is 60 times larger than XLM-RoBERTa$_{\text{base}}$, and when compared to fine-tuning and prompt-tuning methods with limited data ($b = 2$ or 4), our method does not perform better. However, it is important to note that the CLS dataset suffers from class imbalance, where the number of instances belonging to two classes is approximately half of those in the other two classes.

Overall, N2C2 exhibits superior performance over baselines in both accuracy and ECE across all datasets, demonstrating consistent improvement with as the training data increases, although facing challenges with class imbalance in some cases.

### *5.2. Ablation Study*

Table 2 outlines ablation experiments for N2C2 (XLMR-base). In Variant (1), removing the confidence module (4.2) results in a loss of 0.7% and 1% accuracy on MARC and CLS, respectively, along with an increase of 1.3% and 1.1% in ECE. Variant (2) involves $k$NN retrieval in original feature spaces (4.1), leading to a significant 3% accuracy drop for both datasets, yet lower ECE.In Variant (3), fixing retrieval results to the top 8 examples maintains accuracy but notably increases ECE by 5% and 4% on MARC and CLS, respectively. Overall, the confidence-aware distribution construction module can improve accuracy and calibration, while retrieval representation shaping and the weight estimation network (DWE) can aid calibration.

### *5.3. Comparisons With Other Calibration Methods*

Given the effectiveness of N2C2 in reducing ECE, we are interested in examining the performance of classic calibration methods for cross-lingual scenarios. Specifically, we consider the *temperature scaling (TS)* technique [8]

Table 3. Calibration Errors for XLM-RoBERTa$_{\text{base}}$ when using different methods for calibration.

| Dataset | Methods | FT | PT | N2C2 |
|---|---|---|---|---|
| **MARC** | Vanilla | $53.05_{\pm 3.2}$ | $57.94_{\pm 1.3}$ | **$10.89_{\pm 3.5}$** |
| | LS | $45.36_{\pm 5.01}$ | $42.04_{\pm 0.87}$ | |
| | TS | $43.87_{\pm 20.82}$ | $57.94_{\pm 1.41}$ | |
| | LS + TS | $35.93_{\pm 18.0}$ | $42.04_{\pm 0.87}$ | |
| **CLS** | Vanilla | $37.48_{\pm 17.6}$ | $54.06_{\pm 2.3}$ | **$7.61_{\pm 0.6}$** |
| | LS | $29.91_{\pm 18.17}$ | $40.24_{\pm 2.36}$ | |
| | TS | $30.62_{\pm 6.60}$ | $54.06_{\pm 2.69}$ | |
| | LS + TS | $27.17_{\pm 3.17}$ | $40.24_{\pm 2.36}$ | |

and *label smoothing (LS)* [33]. As both methods require training, we solely focus on fine tuning and prompt tuning, and calculate the average results across all languages under 16 shots.

We can observe in Table 3 that both methods significantly reduce the ECE. An exception is the application of temperature scaling during prompt tuning, where its effectiveness is limited. This occurs probably because in prompt tuning (before applying *softmax*) the probabilities of the retrieved labels are extremely small. Despite the effectiveness of these two classical calibration methods, N2C2 consistently outperforms them by a considerable margin. This indicates the robustness and superior performance of N2C2 in achieving enhanced calibration compared to other approaches.

## 6. Conclusion

In our study, we have investigated the performance of multilingual models in cross-lingual ICL scenarios, revealing a notable deficiency in both performance and calibration. To address this, we propose N2C2, a cross-lingual ICL technique that leverages examples from the source language. We conduct experiments on two multilingual sentiment classification datasets, comparing our method with strong baselines and popular calibration methods. The results show N2C2 significantly improves the performance in terms of accuracy and expected calibration errors. Furthermore, our ablation studies demonstrate the contributions of each module within our framework, providing deeper insights into their role and impact. Importantly, our approach exhibits scalability, proving effective even with larger models. Our proposed method not only provides a substantial improvement in cross-lingual ICL, but also offers insights for future research in more effective cross-lingual learning strategies.

## References


[1] K. Ahuja, S. Sitaram, S. Dandapat, and M. Choudhury, "On the calibration of massively multilingual language models," arXiv preprint arXiv:2210.12265, 2022. [Online]. Available: https://arxiv.org/abs/2210.12265.

[2] T. Bose, N. Aletras, I. Illina, and D. Fohr, "Dynamically refined regularization for improving cross-corpora hate speech detection," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 372–382.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, and P. Dhariwal, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., vol. 33, 2020, pp. 1877–1901.

[4] Y. Chen, L. Yuan, G. Cui, Z. Liu, and H. Ji, "A close look into the calibration of pre-trained language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1343–1367.

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Association for Computational Linguistics, Jul. 2020, pp. 8440–8451.

[6] S. Desai and G. Durrett, "Calibration of pre-trained transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Nov. 2020, pp. 295–302.

[7] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, Aug. 2021, pp. 3816–3830.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., vol. 70, Proceedings of Machine Learning Research, PMLR, Aug. 2017, pp. 1321–1330.

[9] T. He, B. McCann, C. Xiong, and E. Hosseini-Asl, "Joint energy-based model training for better calibrated natural language understanding models," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, Apr. 2021, pp. 1754–1761.

[10] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, Proceedings of Machine Learning Research, PMLR, Jun. 2019, pp. 2712–2721.

[11] Y. Hu, C.-H. Lee, T. Xie, T. Yu, N. A. Smith, and M. Ostendorf, "In-context learning for few-shot dialogue state tracking," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 2627–2643.

[12] L. Huang, S. Ma, D. Zhang, F. Wei, and H. Wang, "Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 11488–11497.

[13] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, MIT Press, Cambridge, MA, 2020, pp. 423–438.

[14] Z. Jiang, A. Liu, and B. Van Durme, "Calibrating zero-shot cross-lingual (un-)structured predictions," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 2648–2674.

[15] T. Jung, D. Kang, H. Cheng, L. Mentch, and T. Schaaf, "Posterior calibrated training on sentence classification tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Association for Computational Linguistics, Jul. 2020, pp. 2723–2730.

[16] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith, "The multilingual Amazon reviews corpus," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Nov. 2020, pp. 4563–4568.

[17] S. Kim, D. Ki, Y. Kim, and J. Lee, "Boosting cross-lingual transferability in multilingual models via in-context learning," *ArXiv*, vol. abs/2305.15233, 2023.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2017.

[19] H. Li, H. Yan, Y. Li, L. Qian, Y. He, and L. Gui, "Distinguishability calibration to in-context learning," in *Findings of the Association for Computational Linguistics: EACL 2023*, A. Vlachos and I. Augenstein, Eds. Association for Computational Linguistics, May 2023, pp. 1385–1397.

[20] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, and D. Simig, *et al.*, "Few-shot learning with multilingual generative language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 9019–9052.

[21] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, *et al.*, "What makes good in-context examples for GPT-3?" in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, E. Agirre, M. Apidianaki, and I. Vulić, Eds. Association for Computational Linguistics, May 2022, pp. 100–114.

[22] Y. Liu, T. Schick, and H. Schütze, "Semantic-oriented unlabeled priming for large-scale language models," in *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, N. S. Moosavi, I. Gurevych, Y. Hou, G. Kim, Y. J. Kim, T. Schuster, *et al.*, Eds. Association for Computational Linguistics, Jul. 2023, pp. 32–38.

[23] X. Lyu, S. Min, I. Beltagy, L. Zettlemoyer, and H. Hajishirzi, "Z-ICL: Zero-shot in-context learning with pseudo-demonstrations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, Jul. 2023, pp. 2304–2317.

[24] M. Pakdaman Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Feb. 2015.

[25] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[26] E. Nie, S. Liang, H. Schmid, and H. Schütze, "Cross-lingual retrieval augmented prompt for low-resource languages," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, Jul. 2023, pp. 8320–8340.

[27] F. Nie, M. Chen, Z. Zhang, and X. Cheng, "Improving few-shot performance of language models via nearest neighbor calibration," *arXiv preprint*, arXiv:2212.02216, 2022.

[28] S. Y. Park and C. Caragea, "On the calibration of pre-trained language models using mixup guided by area under the margin and saliency," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, May 2022, pp. 5364–5374.

[29] P. Prettenhofer and B. Stein, "Cross-language text classification using structural correspondence learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, J. Hajič, S. Carberry, S. Clark, and J. Nivre, Eds. Association for Computational Linguistics, Jul. 2010, pp. 1118–1127.

[30] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, Apr. 2021, pp. 255–269.

[31] P. Shi, R. Zhang, H. Bai, and J. Lin, "XRICL: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 5248–5259.

[32] W. Shi, J. Michael, S. Gururangan, and L. Zettlemoyer, "Nearest neighbor zero-shot inference," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 3254–3265.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[34] E. Tanwar, S. Dutta, M. Borthakur, and T. Chakraborty, "Multilingual LLMs are better cross-lingual in-context learners with alignment," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, Jul. 2023, pp. 6292–6307.

[35] J. Wang, C. Wang, F. Luo, C. Tan, M. Qiu, F. Yang, *et al.*, "Towards unified prompt tuning for few-shot text classification," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 524–536.

[36] A. Webson and E. Pavlick, "Do prompt-based models really understand the meaning of their prompts?" in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Association for Computational Linguistics, Jul. 2022, pp. 2300–2344.

[37] G. I. Winata, A. Madotto, Z. Lin, R. Liu, J. Yosinski, and P. Fung, "Language models are few-shot multilingual learners," in *Proceedings of the 1st Workshop on Multilingual Representation Learning*, D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, and G. G. Sahin, Eds. Association for Computational Linguistics, Nov. 2021, pp. 1–15.

[38] BigScience Workshop, T. Le Scao, A. Fan, C. Akiki, E. Pavlick, and S. Ilić, *et al.*, "BLOOM: A 176B-parameter open-access multilingual language model," *arXiv preprint*, arXiv:2211.05100, 2023.

[39] B. Xu, Q. Wang, Z. Mao, Y. Lyu, Q. She, and Y. Zhang, "*k*NN prompting: Beyond-context learning with calibration-free nearest neighbor inference," in *Proceedings of The Eleventh International Conference on Learning Representations*, 2023.

[40] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, *et al.*, Eds. Association for Computational Linguistics, Jun. 2021, pp. 483–498.

[41] M. Zhao and H. Schütze, "Discrete and soft prompting for multilingual models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Association for Computational Linguistics, Nov. 2021, pp. 8547–8555.

[42] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds. PMLR, Jul. 2021, vol. 139, pp. 12697–12706.

[43] X. Zheng, Z. Zhang, J. Guo, S. Huang, B. Chen, W. Luo, *et al.*, "Adaptive nearest neighbor machine translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, Aug. 2021, pp. 368–374.

[44] W. Long and B. Webber, "Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations," *arXiv preprint arXiv:2301.02724*, Jan. 2023. [Online]. Available: https://arxiv.org/abs/2301.02724

[45] W. Long, B. Webber, and D. Xiong, "TED-CDB: A large-scale Chinese discourse relation dataset on TED talks," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Online, Nov. 2020, pp. 2793–2803. [Online]. Available: https://aclanthology.org/2020.emnlp-main.223/

[46] W. Long, X. Cai, J. Reid, B. Webber, and D. Xiong, "Shallow discourse annotation for Chinese TED talks," in *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, Marseille, France, May 2020, pp. 1025–1032. [Online]. Available: https://aclanthology.org/2020.lrec-1.129/

[47] W. Long, S. N, and B. Webber, "Multi-label classification for implicit discourse relation recognition," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, Bangkok, Thailand, Aug. 2024, pp. 8437–8451. [Online]. Available: https://aclanthology.org/2024.findings-acl.500/

[48] J. He, W. Long, and D. Xiong, "Evaluating discourse cohesion in pre-trained language models," in *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, C. Braud, C. Hardmeier, J. J. Li, S. Loaiciga, M. Strube, and A. Zeldes, Eds. International Conference on Computational Linguistics, Gyeongju, Republic of Korea and Online, Oct. 2022, pp. 28–34. [Online]. Available: https://aclanthology.org/2022.codi-1.4/

[49] W. Long and B. Webber, "Leveraging hierarchical prototypes as the verbalizer for implicit discourse relation recognition," *arXiv preprint arXiv:2411.14880*, Nov. 2024. [Online]. Available: https://arxiv.org/abs/2411.14880

[50] J. He, Y. Yang, W. Long, D. Xiong, V. G. Basulto, and J. Z. Pan, "Evaluating and improving graph to text generation with large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Association for Computational Linguistics, Albuquerque, New Mexico, Apr. 2025, pp. 10219–10244. [Online]. Available: https://aclanthology.org/2025.naacl-long.513/

[51] J. He, N. Hu, W. Long, J. Chen, and J. Z. Pan, "MINTQA: A multi-hop question answering benchmark for evaluating LLMs on new and tail knowledge," *arXiv preprint arXiv:2412.17032*, Dec. 2025. [Online]. Available: https://arxiv.org/abs/2412.17032

[52] A. Keleg, M. Lindemann, D. Liu, W. Long, and B. L. Webber, "Automatically discarding straplines to improve data quality for abstractive news summarization," in *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, T. Shavrina, V. Mikhailov, V. Malykh, E. Artemova, O. Serikov, and V. Protasov, Eds. Association for Computational Linguistics, Dublin, Ireland, May 2022, pp. 42–51. [Online]. Available: h