Steering LLMs toward Korean Local Speech: Iterative Refinement Framework for Faithful Dialect Translation

Keunhyeung Park, Seunguk Yu, Youngbin Kim
Chung-Ang University, Seoul, Republic of Korea
synoark99@cau.ac.kr, seungukyu@gmail.com, ybkim85@cau.ac.kr

Abstract

Standard-to-dialect machine translation remains challenging due to a persistent dialect gap in large language models and evaluation distortions inherent in n-gram metrics, which favor source copying over authentic dialect translation. In this paper, we propose the dialect refinement (**DIA-REFINE**) framework, which guides LLMs toward faithful target dialect outputs through an iterative loop of translation, verification, and feedback using external dialect classifiers. To address the limitations of n-gram-based metrics, we introduce the dialect fidelity score (**DFS**) to quantify linguistic shift and the target dialect ratio (**TDR**) to measure the success of dialect translation. Experiments on Korean dialects across zero-shot and in-context learning baselines demonstrate that DIA-REFINE consistently enhances dialect fidelity. The proposed metrics distinguish between **False Success** cases, where high n-gram scores obscure failures in dialectal translation, and **True Attempt** cases, where genuine attempts at dialectal translation yield low n-gram scores. We also observed that models exhibit varying degrees of responsiveness to the framework, and that integrating in-context examples further improves the translation of dialectal expressions. Our work establishes a robust framework for goal-directed, inclusive dialect translation, providing both rigorous evaluation and critical insights into model performance.

Keywords: Korean Dialect Translation, Iterative Refinement, Dialectal Evaluation Bias

1. Introduction

Recent advancements in large language models (LLMs) have revolutionized the field of machine translation (Lyu et al., 2024), highlighting the importance of inclusive approaches for diverse low-resource languages (Costa-jussà et al., 2024). In this context, dialect research holds significant academic and social value, as it contributes to preserving the linguistic heritage of potentially marginalized regional languages and enhancing technological inclusivity (Ziems et al., 2022).

Despite the progress of LLMs, a significant performance disparity persists between standard and non-standard dialects (Faisal et al., 2024; Fleisig et al., 2024; Kantharuban et al., 2023). This behavior exhibits an asymmetry depending on the translation direction, with translating from a standard language to a dialect being more challenging than the reverse (Park et al., 2020; Zheng et al., 2022). The difficulty of LLMs in translating distinctive dialectal features is evidenced in our experimental results, where some models achieved a zero-shot success rate of only 2%, with most outputs closely resembling the standard language. While dialect machine translation exhibits these limitations, research on controlling and evaluating the dialectal fidelity of translation outputs remains scarce.

To address this gap, we propose a framework that utilizes external feedback from dialect classifiers to guide LLMs toward translating target dialect outputs, as illustrated in Figure 1. Our core methodology, dialect refinement (**DIA-REFINE**), verifies whether the output possesses linguistic features of the target dialect based on the dialect classifier's prediction. If the output fails to satisfy the target dialect condition, explicit feedback derived from the classification result is provided to the model for re-translation. Through this iterative process, the LLM can be effectively controlled to perform goal-oriented dialect translation.

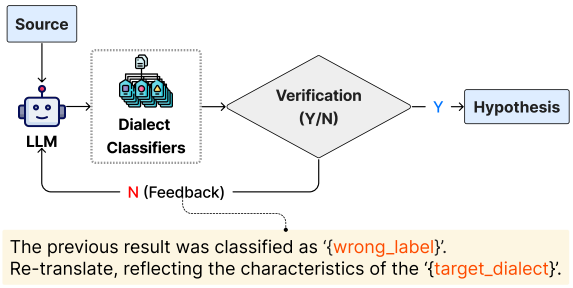

Figure 1: An overview of our dialect refinement (DIA-REFINE) framework. The LLM's output is verified by an external ensemble of dialect classifiers, which provides explicit feedback to guide the model toward translating the target dialect.

However, traditional n-gram-based metrics such as BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) have been repeatedly reported to have a structural blind spot (Aepli et al., 2023; Chen et al., 2024). This vulnerability can lead to higher scores for **False Success**, which involves merely copying the source, than for a

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3257–3269
11–16 May 2026.

Variants	Forms	Analysis / Key Features	BLEU	chrF++	**DFS**
Source (Standard)	제주도는 특별한 약초들이 많이 없으니까 그냥 그냥 넘긴 거 같아	Standard grammar and vocabulary.	–	–	–
Reference (*Jeju* Dialect)	제주도는 특별한 약초들이 많이 엇이난 그냥 그냥 넘긴 거 닮아	Contains target dialectal features.	–	–	–
Hypothesis 1	제주도엔 특별한 약초덜이 많이 읎응께 그냥 그냥 넘긴 거 같아	**True Attempt**: Generates multiple dialectal features (-엔 (-en), -덜 (-deol), -읎응께 (-eup-eung-kke))	27.78	36.80	**0.672**
Hypothesis 2	제주도는 특별한 약초들이 많이 없으니깐 그냥 그냥 넘긴 거 같아	**False Success**: Output mirrors the standard source rather than the intended dialect.	**52.54**	**67.82**	-0.672

Table 1: Evaluation distortions in Korean dialect translation: while n-gram metrics favor surface overlap (**False Success**), the proposed DFS metric accurately captures the shift toward the dialectal output (**True Attempt**). Underlined morphemes denote dialectal expressions in the hypothesis. The shared meaning is '*I think they just passed on Jeju Island since it doesn't have a lot of special medicinal herbs there.*'

True Attempt to generate dialectal features[1]. As shown in Table 1, this paradox causes outputs that merely copy the standard source to be evaluated more favorably than genuine attempts to generate dialectal features. This biased evaluation system hinders the accurate assessment of practical gains of the dialectal translation outputs.

To address the limitations of existing metrics and rigorously validate our framework, we introduce a new multifaceted evaluation system. It comprises the dialect fidelity score (**DFS**), which quantifies the linguistic shift of translated outputs toward the dialectal reference, and the target dialect ratio (**TDR**), which measures the proportion of outputs correctly classified as the target dialect. These metrics are designed to reliably assess the effectiveness of our framework in steering translation toward the target dialect. To validate the proposed framework, we conducted experiments translating standard Korean into three major dialects with distinct segmental features, *Jeolla*, *Gyeongsang*, and *Jeju*, two of which have been largely unaddressed in previous research. Using several state-of-the-art LLMs, we performed a comparative analysis of existing approaches, including zero-shot and in-context learning, against our proposed `DIA-REFINE` framework to evaluate model performance under various conditions. The main contributions of this study are as follows:

(i) We propose the dialect refinement (**`DIA-REFINE`**) framework, an effective mechanism for controlling dialect translation.

(ii) We introduce a robust evaluation system with **DFS** and **TDR** metrics, addressing the limitations of n-gram-based metrics and enabling an accurate assessment of dialectal fidelity.

(iii) Our analysis demonstrates that the proposed metrics effectively distinguish between cases of **False Success** and **True Attempt** across models. Moreover, we establish that combining our framework with in-context examples yields the most effective performance in dialect translation.

This paper is organized as follows: Section 2 reviews related work on recent trends in dialect research, steering generation and challenges in dialect evaluation. Section 3 and 4 present the `DIA-REFINE` framework and experimental setup, respectively. Section 5 details our evaluation metrics. Section 6 reports the main results and our detailed analysis, and Section 7 concludes. We plan to release the code to facilitate the proposed framework and metrics, thereby supporting further research in this area[2].

2. Related Work

2.1. Recent Trends in Dialect Research

NLP research has traditionally focused on a small number of standard, high-resource languages (Nigatu et al., 2024; Ziems et al., 2022). This has created a significant performance disparity between standard and non-standard varieties. For

[1] In this paper, we define **False Success** as cases exhibiting relatively high n-gram but low DFS/TDR scores, and **True Attempt** as cases showing the opposite pattern, with low n-gram and high DFS/TDR scores, within our comparative analysis across models.

[2] Due to the source dataset's terms of use, the processed dataset cannot be released; however, we made our code publicly available to support reproducibility. Available at: https://github.com/keunhyeung/DIA-REFINE

instance, tweets in African American English are up to twice as likely to be mislabeled as offensive (Sap et al., 2019), and models like ChatGPT produce responses to non-standard dialects that are perceived as more stereotypical and demeaning (Fleisig et al., 2024). These biases perpetuate linguistic discrimination and marginalize speakers of non-standard dialects.

To address these limitations, a wider variety of dialects has received increasing attention in recent research. These efforts have led to the creation of notable resources, such as the parallel data MADAR for Arabic (Bouamor et al., 2018) and YORÙLECT for Yorùbá (Ahia et al., 2024). ARGEN benchmark offers a comprehensive framework for evaluating dialect generation (Nagoudi et al., 2022). The latest benchmark WMT24++ includes 10 dialects across 5 languages (Deutsch et al., 2025). Collectively, these efforts reflect a trend toward addressing a broader range of language varieties and underscore the need for dialect-specific research. In contrast to this trend, research on Korean has been limited, focusing almost exclusively on the *Jeju* dialect through efforts to construct the JIT dataset (Park et al., 2020) and cross-lingual pretraining with Japanese (Zheng et al., 2022). Our work is significant in that it expands this scope to include the *Jeolla* and *Gyeongsang* dialects, which have been largely unaddressed in previous Korean dialect research.

2.2. Generation Control using External Feedback

Leveraging external information to control model outputs toward specific goals is an active line of research. Classifier-guided diffusion steers the sampling process through the gradient of a pre-trained classifier, thereby improving the quality of the generated samples (Dhariwal and Nichol, 2021). SELF-REFINE iterates a generate–feedback–refine loop with a single model and requires no supervised data (Madaan et al., 2023). In machine translation tasks, iteratively prompting an LLM to refine its result can improve the fluency and naturalness itself, mimicking a human-like editing process (Chen et al., 2024).

In this paper, we extend the idea of iterative refinement by using external feedback to control LLM output quality for the dialect translation task. In our DIA-REFINE framework, the dialect classifier functions as a linguistic tool to verify the dialectal fidelity of the model's output. We explicitly utilized the feedback from the high-performance ensemble model for a task-specific control that guides the dialect translation process.

Class	Precision	Recall	F1
Jeolla	0.9308	0.9150	0.9228
Gyeongsang	0.9135	0.9080	0.9107
Jeju	0.9888	0.9720	0.9803
Standard	0.9188	0.9510	0.9346
Unknown	0.9950	1.0000	0.9975

Overall Metric	Score
Accuracy	0.9492
Macro F1-Score	0.9494

Table 2: Performance of the best ensemble dialect classifier from our experiments, with the top showing per-class results and the bottom table showing the overall average.

2.3. Distortion in Dialect Translation Evaluation

Despite the widespread adoption of embedding-based metrics for evaluating machine translation tasks (Larionov et al., 2024), these advanced metrics show a significant lack of dialect robustness (Sun et al., 2023). Therefore, research on standard-to-dialect translation has still heavily relied on n-gram metrics such as BLEU and chrF (Ahia et al., 2024; Faisal et al., 2024; Nagoudi et al., 2022; Zheng et al., 2022).

However, these metrics can distort evaluation by overweighting surface-form overlap rather than translation quality. Aepli et al. (2023) showed that for Swiss German, a dialect without standardized orthography, semantically incorrect yet lexically similar outputs often scored higher on n-gram metrics than correct ones. In this aspect, Chen et al. (2024) report that, in multilingual translation tasks, iterative translation refinement tends to reduce n-gram–based scores even as human preferences increase, revealing a divergence between n-gram–based and semantic evaluations.

This **False Success** phenomenon has been reported across dialect studies. Park et al. (2020) found that a trivial copy-model achieved high BLEU scores on translation from standard Korean to the *Jeju* dialect. Similarly, Liu (2022) found that BLEU can systematically favor copy-bias models in Cantonese translation. Analyzing multi-dialect performance, Kantharuban et al. (2023) quantified a strong correlation between dialect–standard lexical similarity and performance on n-gram metrics. Collectively, these studies show that relying solely on n-gram metrics for standard-to-dialect evaluation can severely distort assessment. To address this evaluation gap, we propose two complementary metrics, DFS and TDR, to enable a more accurate measure of dialectal fidelity.

Method	Candidates	Selection	Verification	Feedback Loop
Baseline (ZS, ICL)	1	N/A	N/A	N/A
Baseline + DIA-REFINE(S)	1	N/A	$\mathcal{C}(y) = d_{\text{tgt}}$	Retry on mismatch
Baseline + DIA-REFINE(M)	$k \in \{3, 4, 5\}$	Select top-1 as y^*	$\mathcal{C}(y^*) = d_{\text{tgt}}$	Retry on mismatch

Table 3: Variants of the proposed DIA-REFINE differing in candidate generation and hypothesis selection strategies. The framework extends the baseline with verification and feedback loops, allowing up to two retries. For DIA-REFINE(M), the best candidate y^* is selected from k options by maximizing their posterior probability. Throughout this process, we employed our trained ensemble dialect classifier $\mathcal{C}$.

3. Proposed Methodology: The DIA-REFINE Framework

We propose dialect-refinement (**DIA-REFINE**), a framework designed to steer LLMs toward consistent and high-fidelity dialect translation. At its core, DIA-REFINE operates through an iterative loop of translation and verification, using feedback from external dialect classifiers to progressively guide the translation output toward the target dialect.

3.1. Dialect Classifier Dataset Construction

We collected the Korean dialect data corpus (National Information Society Agency (NIA), 2022), a public resource to preserve endangered dialects. It is a parallel corpus of standard Korean and dialectal sentence pairs from five major regions: *Gangwon*, *Chungcheong*, *Jeolla*, *Gyeongsang*, and *Jeju*.

From this corpus, we constructed a dialect classifier dataset. We selected *Jeolla*, *Gyeongsang*, and *Jeju* as the target dialects due to their salient features[3]. These were supplemented by *Standard* and *Unknown* classes, representing no dialect and mixed-dialect sentences. We compiled 10,000 samples per class to prioritize robustness, resulting in 50,000 total samples. We divided each class into a 9:1 ratio for training and evaluation samples. Class-specific curation is as follows:

- **The *Dialect* classes** (*Jeolla*, *Gyeongsang*, and *Jeju*) include only samples where the normalized Levenshtein distance (Yujian and Bo, 2007) between the standard source and its dialect counterpart is ≥ 0.1, ensuring form-level divergence (Johannessen et al., 2020).
- **The *Standard* class** comprises no dialect samples drawn at random from the source.
- **The *Unknown* class** is constructed as a hard-negative set to make the dialect classifier more robust (Li et al., 2024). We first extracted salient dialect lexical features using TF-IDF and then prompted Gemini-2.5-Flash-Lite[4] to generate sentences mixing distinct dialects that do not conform to a single dialect. This process was conducted to prevent overfitting and avoid ambiguous outputs that LLMs may produce.

3.2. Building an Ensemble of Dialect Classifiers

We adopted an ensemble approach (Arango et al., 2024) to enhance robustness and stability of the dialect classifier $\mathcal{C}$. We fine-tuned five distinct Korean pre-trained language models[5][6][7][8][9] using a set of shared hyperparameters, with a batch size of 16, a learning rate of 4e-5, a weight decay of 0.01, 4 training epochs, a maximum input length of 64, and a seed of 1337 for reproducibility.

After fine-tuning each model, we searched for the optimal ensemble scenario among the 31 combinations. While the lowest-performing case yielded an accuracy of 92.92%, we found that BM-K/KoSimCSE-roberta[7] combined with beomi/KcELECTRA-base[9] performed best. As shown in Table 2, this ensemble achieved an overall accuracy of 94.92% with balanced performance.

3.3. DIA-REFINE Framework for Dialect Translation

We operate our **DIA-REFINE** on the trained ensemble dialect classifier. First, an LLM generates a

[3]The *Jeolla* and *Gyeongsang* dialects are clearly distinct from standard Korean, while the *Jeju* dialect has a unique feature owing to its geographically isolated location. In contrast, the *Gangwon* and *Chungcheong* dialects are generally considered variants of standard Korean, exhibiting fewer linguistic differences.

[4]https://aistudio.google.com/

[5]https://huggingface.co/klue/bert-base

[6]https://huggingface.co/klue/roberta-large

[7]https://huggingface.co/BM-K/KoSimCSE-roberta

[8]https://huggingface.co/monologg/koelectra-base-v3-discriminator

[9]https://huggingface.co/beomi/KcELECTRA-base

candidate translation y following the given instruction, then $\mathcal{C}$ determines whether the y aligns with the intended target dialect d_{tgt}. If it was classified as a non-targeted dialect (i.e., $\mathcal{C}(y) \neq d_{tgt}$), the framework requests retranslation using a prompt that contains explicit feedback. This feedback goes beyond a simple mismatch signal, specifying which class the output was misclassified into to guide the correction (Dhariwal and Nichol, 2021).

Furthermore, when the DIA-REFINE detects the model oscillating between different error types on consecutive retries, the prompt explicitly flags this pattern to help the model break the cycle and steer its generation more effectively. As detailed in Table 3, we defined two variants of our framework that primarily differ in their candidate generation and hypothesis selection strategies:

- **DIA-REFINE(S) (Single-candidate)** follows a simple single-candidate approach that verifies each output as it is generated.
- **DIA-REFINE(M) (Multi-candidates)** generates k outputs and selects the most promising candidate y^* based on the dialect classifier's posterior probability. Only this selected candidate is then verified. This generate-and-select approach aims to broaden the search space, with k starting at 3 and incrementing by one with each retry to enhance exploration.

Both variants allow up to two retries on mismatch, for a total of three attempts.

4. Experimental Setup

4.1. Dialect Translation Test Set

From the corpus in Section 3.1, we created a dialect translation test set to evaluate translation performance. For this set, we randomly sampled 300 pairs each from the *Jeolla*, *Gyeongsang*, and *Jeju* dialects, which exhibit salient features. We only included sentences with a length of ≥ 30 characters to ensure they contained sufficient content.

4.2. Models and Baselines

Given our focus on Korean dialects, our evaluation includes the three Korean-specialized models HyperCLOVAX[10], EXAONE-3.5[11], and EEVE[12]. To provide a broader comparative context, we also evaluated an open general-purpose model Llama-3.1[13], and a proprietary model Gemini-1.5[4].

We included baselines with zero-shot (ZS) and in-context learning (ICL). The latter is known to be effective in low-resource translation (Pei et al., 2025). From the corpus in Section 3.1, we prepared an ICL example pool of 10,000 pairs for each class, ensuring no overlap with the test set. We retrieved the top-10 most relevant pairs from the example pool using BM25 to construct ICL examples.

Model	ZS		ICL+DIA-REFINE(M)	
	Vanilla	Fine-tuned	Vanilla	Fine-tuned
HyperCLOVAX	−0.01	−0.37	0.00	−0.06
Llama-3.1	−0.01	−0.37	0.00	−0.10
EEVE	−0.01	−0.38	0.00	−0.05
EXAONE-3.5	−0.01	−0.39	0.00	0.27
Gemini-1.5	0.00	0.36	0.00	0.39

Table 4: Averaged DFS scores between vanilla model[9] and fine-tuned variants, evaluated on the dialect translation test set. The dialect-aware embeddings yield pronounced scores with sensitivity to dialectal features, whereas the standard embeddings indicate reduced effectiveness.

5. Evaluation Framework

5.1. Limitations of N-gram Metrics

By overweighting surface-form similarity rather than dialect fidelity, n-gram metrics can award high scores to outputs to the standard form, yielding **False Success** (Liu, 2022; Park et al., 2020). For instance, Hypothesis 2 in Table 1, which largely mirrors the standard source, achieves deceptively high scores of 52.54 BLEU and 67.82 chrF++. In contrast, Hypothesis 1 makes **True Attempt** by introducing dialect features like -엔 (-en), -덜 (-deol), and -읍응께 (-eup-eung-kke). Although these features do not belong to the target *Jeju* dialect, this genuine effort is penalized with significantly lower scores of 27.78 BLEU and 36.80 chrF++. This example demonstrates that relying solely on n-gram metrics is not only insufficient but also potentially misleading, thereby making them unsuitable for properly validating the effectiveness of the DIA-REFINE framework.

5.2. Proposed Metrics: DFS and TDR

To address this, we complement n-gram metrics with our two proposed metrics. First, **dialect fidelity score (DFS)** is a continuous measure of whether the hypothesis h is linguistically closer to the dialectal reference r than to the standard source s. In

[10] https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-1.5B

[11] https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct

[12] https://huggingface.co/yanolja/YanoljaNEXT-EEVE-Instruct-10.8B

[13] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Metric	Method	HyperCLOVAX	Llama-3.1	EEVE	EXAONE-3.5	Gemini-1.5	Model Avg
chrF++	ZS	60.33	68.97	54.03	29.05	39.50	50.38
	ZS + DIA-REFINE(S)	60.76	62.52	52.41	29.94	38.03	48.73
	ZS + DIA-REFINE(M)	52.52	44.29	39.82	25.20	38.55	40.08
	ICL	59.17	71.31	70.99	65.41	63.08	65.99
	ICL + DIA-REFINE(S)	55.35	70.42	70.48	60.19	62.13	63.72
	ICL + DIA-REFINE(M)	46.00	66.15	65.67	54.95	61.68	58.89
BLEU	ZS	47.64	58.53	41.71	11.44	23.06	36.48
	ZS + DIA-REFINE(S)	47.94	49.82	39.49	11.97	21.92	34.23
	ZS + DIA-REFINE(M)	38.82	28.98	25.24	8.13	22.13	24.66
	ICL	46.21	61.75	61.34	52.88	49.10	54.25
	ICL + DIA-REFINE(S)	41.44	60.57	60.68	45.75	47.73	51.23
	ICL + DIA-REFINE(M)	31.70	54.67	54.30	39.14	47.22	45.41
DFS	ZS	-0.37	-0.37	-0.38	-0.39	0.36	-0.23
	ZS + DIA-REFINE(S)	-0.37	-0.37	-0.37	-0.32	0.38	-0.21
	ZS + DIA-REFINE(M)	-0.32	-0.32	-0.28	-0.08	0.40	-0.12
	ICL	-0.30	-0.26	-0.26	-0.04	0.31	-0.11
	ICL + DIA-REFINE(S)	-0.24	-0.24	-0.23	0.07	0.36	-0.06
	ICL + DIA-REFINE(M)	-0.06	-0.10	-0.05	0.27	0.39	0.09
TDR	ZS	0.03	0.03	0.02	0.01	0.93	0.20
	ZS + DIA-REFINE(S)	0.04	0.03	0.03	0.08	0.98	0.23
	ZS + DIA-REFINE(M)	0.08	0.08	0.13	0.42	0.99	0.34
	ICL	0.09	0.15	0.16	0.41	0.85	0.33
	ICL + DIA-REFINE(S)	0.15	0.18	0.19	0.60	0.97	0.42
	ICL + DIA-REFINE(M)	0.41	0.36	0.42	0.94	0.99	0.62

Table 5: Performance comparison of five language models across six methodologies, evaluated with four key metrics. All scores are averaged over three target dialects (*Jeolla*, *Gyeongsang* and *Jeju*). The Model Avg column presents the average performance across all models for each method.

an embedding space, we compute the log ratio of their cosine similarities, as shown in Eq. (1). We add 1 to the cosine similarity and a small constant ε $(= 10^{-6})$ to ensure numerical stability for the logarithm. A positive DFS indicates a shift toward the dialect, while a negative value indicates proximity to the standard variety.

$$\mathrm{DFS} = \log\left(\frac{1 + \cos(\mathbf{e}_h, \mathbf{e}_r) + \varepsilon}{1 + \cos(\mathbf{e}_h, \mathbf{e}_s) + \varepsilon}\right), \qquad (1)$$

To ensure reliable representations, we extracted dialect-aware embeddings $\mathbf{e}_h, \mathbf{e}_r, \mathbf{e}_s$ using our fine-tuned `beomi/KcELECTRA-base` model, with the final classification layer removed. As shown in Table 4, embeddings from the dialect-aware model yielded reasonable DFS scores compared to the vanilla model. While our model exhibits a score range from -0.39 to 0.27 across methods, the vanilla model shows lower sensitivity, varying only from -0.01 to 0. This demonstrates that our dialect-aware model yields an embedding space that truly reflects the intended dialectal features.

The effectiveness of this metric is apparent when returning to the examples in Table 1. The **True Attempt** of Hypothesis 1, despite its low n-gram scores, yields a high positive DFS of 0.672, correctly capturing its linguistic deviation from the standard source. Conversely, the **False Success** of Hypothesis 2 yields a negative DFS of -0.672, indicating its close alignment with the standard source. In our low-resource language setting using dialects, where embedding-based metrics are previously constrained (Mukherjee et al., 2025), DFS highlights the potential of an embedding-based evaluation. However, it is important to note that DFS is not a standalone measure of overall translation success and should therefore be interpreted in conjunction with other metrics.

Second, **target dialect ratio (TDR)** represents the proportion of outputs correctly generated in the target dialect. For a set of generated hypotheses H, it is the proportion of the hypotheses classified by $\mathcal{C}$ as the target dialect d_{tgt}, as shown in Eq. (2). This metric provides a clear quantification of the

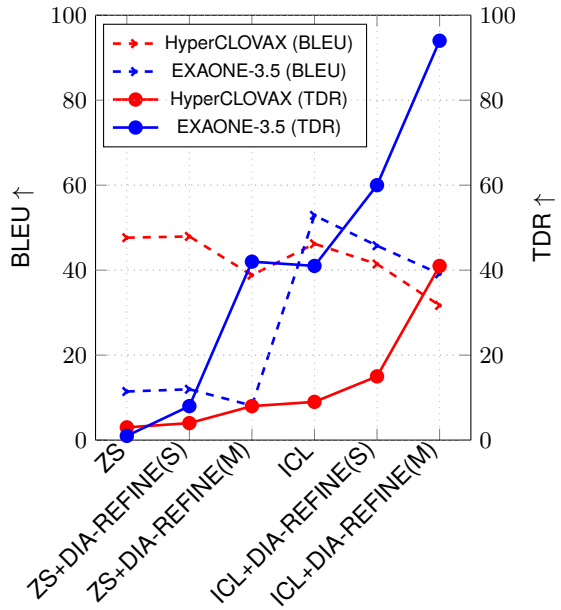

Figure 2: Divergent trends between BLEU and the proposed TDR metric. This highlights a key limitation of BLEU, as it fails to reflect the positive impact of our methods on dialect translation.

actual success rate in dialect translation.

$$\text{TDR} = \frac{|\{y \in H \mid \mathcal{C}(y) = d_{\text{tgt}}\}|}{|H|}. \quad (2)$$

We used DFS and TDR to mitigate the limitations of n-gram-based metrics and the absence of reliable dialectal embedding-based metrics.

6. Results and Analysis

6.1. Unmasking False Success Cases

Our results clearly demonstrate the limitations of n-gram metrics and validate the necessity of our proposed metrics. As detailed in Table 5, in the ZS setting, several models achieved deceptively high n-gram scores despite poor dialect translation. HyperCLOVAX, Llama-3.1, and EEVE achieved high chrF++/BLEU scores of 60.33/47.64, 68.97/58.53, and 54.03/41.71, respectively. However, these scores significantly contrast with their performance on the proposed metrics, with TDRs all $\leq$ 0.03 and consistently negative DFS scores around -0.37. This indicates that they defaulted to the standard language in nearly 97% of cases.

As shown in Figure 2, the dashed and solid lines represent the BLEU and TDR scores, corresponding to the method used. While the use of ICL and `DIA-REFINE(M)` led to a steep increase in TDR scores, the BLEU scores did not show an evident correlation. If BLEU had been sensitive to the changes in the dialectal output brought by each method, the dashed line would have mirrored that of the solid line, at least for a highly effective translation method like ICL.

A clear illustration of this evaluation paradox is provided by Gemini-1.5, which achieved comparatively low chrF++/BLEU scores of 39.50/23.06 in the ZS setting, yet exhibited a notably high TDR of 0.93. Relying solely on n-gram metrics would have mischaracterized Gemini-1.5 as the lowest-performing model, when in fact it was the most successful model. This finding underscores the limitations of n-gram-based traditional metrics and suggests that employing the proposed DFS and TDR offers a more accurate and reliable evaluation of dialectal outputs by revealing the **False Success** phenomenon in our experiments.

6.2. Performance Trajectory across Methodologies

As detailed in Table 5, the models exhibited distinct performance trajectories. Gemini-1.5, in particular, demonstrated a pronounced performance trajectory, achieving high fidelity to dialectal forms in the zero-shot setting and enhancing translation quality with the relevant examples. This exceptional capability likely stems from its massive scale, affording better instruction adherence and broader linguistic knowledge (Brown et al., 2020; Chung et al., 2024). In the ZS setting with `DIA-REFINE(M)`, the model achieved a DFS of 0.40 and almost perfect TDR of 0.99. However, its n-gram scores were low with chrF++/BLEU scores of 38.55/22.13, indicating a prioritization of dialectal form over lexical overlap with the reference. The introduction of ICL led to a substantial improvement in translation quality, and with `DIA-REFINE(M)`, it yielded chrF++/BLEU scores of 61.68/47.22 while maintaining the exceptional TDR of 0.99. This trajectory suggests that in the zero-shot setting, the model initially focused on satisfying the formal requirements of the dialect classifier and subsequently enhanced translation accuracy by leveraging in-context examples.

Our experimental results also revealed that models exhibit varying degrees of responsiveness across the different methods. HyperCLOVAX, Llama-3.1, and EEVE predominantly exhibited a **False Success** tendency. They demonstrated limited improvement, achieving high n-gram scores but negative DFS scores and TDRs of around 0.40, even when ICL and `DIA-REFINE(M)` were applied simultaneously.

In contrast, EXAONE-3.5 initially presented a clear case of failure, recording low scores on both n-gram and the proposed metrics. This model, however, exhibited a remarkable improvement following the integration of ICL and `DIA-REFINE(M)`, with its DFS increasing to 0.27 and TDR to 0.94,

Method	chrF++	BLEU	**DFS**	**TDR**	**Attempts**
ZS	63.34	51.62	-0.37	0.00	1.00
ZS + DIA-REFINE(S)	60.49	47.59	-0.36	0.01	2.99
ZS + DIA-REFINE(M)	46.44	32.16	-0.29	0.09	2.95
ICL	74.35	64.74	-0.34	0.00	1.00
ICL + DIA-REFINE(S)	70.99	60.48	-0.30	0.06	2.95
ICL + DIA-REFINE(M)	62.64	50.59	**-0.12**	**0.28**	**2.81**

Table 6: Performance comparison on a shared subset of 2,305 difficult samples based on the averaged results across five models. The multi-candidate variant, DIA-REFINE(M), demonstrates the most pronounced corrective effect in converting initial failures into successful dialect translations.

along with a corresponding improvement in n-gram scores. These results suggest that the effectiveness of such approaches for dialect translation may rely on the model's inherent linguistic capability.

6.3. Comparison of DIA-REFINE Variants under Challenging Conditions

To evaluate dialect translation under more challenging conditions, we compiled a separate set of difficult samples. Specifically, during the experiments reported in Table 5, we collected all instances in which every DIA-REFINE method failed on the first attempt, yielding a total of 2,305 samples[14]. The results of experiments conducted exclusively on this subset are presented in Table 6. Under both the ZS and ICL settings, the application of DIA-REFINE(S) and (M) yielded incremental improvements in the proposed DFS and TDR metrics. In particular, TDR increased to 0.09 under the ZS and to 0.28 under ICL setting, demonstrating that relevant examples remain highly influential even for these challenging cases.

It is noteworthy that, even when only the more challenging samples were selected, the proposed metrics exhibited lower scores due to its difficulty, whereas the n-gram-based metrics such as chrF++ and BLEU remained relatively stable. This observation suggests that the high n-gram scores likely stem from models replicating the source sentences rather than accurately generating the intended dialectal translations. Such results underscore the limitations of traditional metrics in faithfully evaluating dialect translation.

In summary, these results demonstrate the superior corrective capability of DIA-REFINE(M). The framework not only achieves a markedly higher success rate but also exhibits enhanced efficiency. In the ICL setting, it attained the highest TDR with an average of 2.81 attempts, compared to 2.95 attempts for DIA-REFINE(S). These findings indicate that the multi-candidate strategy constitutes the most effective mechanism for recovering from initial failures and achieving successful dialect translations under challenging conditions.

[14]We selected 2,305 samples from an initial pool of 4,500 test instances, generated by evaluating 5 models on 300 samples for each of 3 dialects.

7. Conclusion

In this study, we addressed the challenges of dialect translation and evaluation in LLMs. We introduced dialect refinement (**DIA-REFINE**), a framework that steers models toward producing high fidelity dialect translations by leveraging feedback from our ensemble dialect classifiers. Through an iterative process of translation, verification, and feedback, the proposed framework improves dialect translation success rates and fidelity, with the multi-candidate variant proving to be the most effective. Furthermore, to overcome the limitations of traditional evaluation, we proposed new metrics comprising the DFS and TDR. We demonstrated that these metrics can identify cases of **False Success**, where n-gram-based scores reward models for defaulting to the standard language, thereby enabling a more faithful evaluation of true dialect translation capability.

Our analysis revealed that the effectiveness of the DIA-REFINE framework depends on the intrinsic capability of the base model. We found that models exhibiting a **False Success** tendency were less responsive to the proposed framework than those exhibiting **True Attempt**, and even the model that initially failed showed substantial improvement with the DIA-REFINE(M). This study not only presents a robust framework for enhancing dialect translation but also establishes a reliable evaluation that facilitates deeper insights into model behavior beyond conventional metrics.

8. Ethical considerations

Our study utilizes the 'Korean dialect data of middle-aged and elderly speakers' corpus (National Information Society Agency (NIA), 2022), which may introduce potential bias by not fully representing the linguistic diversity across age groups. Given the use of LLMs, there exist inherent concerns regarding model biases. To mitigate this, we strictly limited their application to dialect translation and did not use them for any other purpose. This paper used datasets from The Open AI Dataset Project (AI-Hub, S. Korea). All data information can be accessed through *AI-Hub (www.aihub.or.kr).

9. Limitations

As this study focuses on a single language and a corpus filtered for longer sentences, its generalizability is constrained; however, we expect that the proposed framework can be readily extended to other languages and corpora. Given that we focused on dialects with pronounced features, additional methods are required to enable models to handle dialects lacking distinctive characteristics. The efficacy of the framework depends on an external classifier that requires labeled data, leaving its performance under data-scarce conditions undetermined. Semantic preservation was not evaluated, which is a common limitation in low-resource settings. Our evaluation relies on automatic metrics, while qualitative aspects such as fluency and naturalness require verification by native speakers, which we plan to address in future work.

10. Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

11. Bibliographical References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation (WMT'23)*, pages 1045–1065, Singapore. Association for Computational Linguistics (ACL).

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 4392–4409, Miami, USA. Association for Computational Linguistics (ACL).

Sebastian Pineda Arango, Maciej Janowski, Lennart Purucker, Arber Zela, Frank Hutter, and Josif Grabocka. 2024. Ensembling finetuned language models for text classification. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning (FITML'24)*, Vancouver, Canada. Conference on Neural Information Processing Systems (NeurIPS).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 3387–3396, Miyazaki, Japan. European Language Resources Association (ELRA).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS'20)*, pages 1877–1901. Conference on Neural Information Processing Systems (NeurIPS).

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (EAMT'24)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen,

Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Susan Yeom, Bhavick Rungta, Juan Pino, Francisco Guzmán, Hassan Sajjad, Simran Khanuja, Madian Khabsa, Jesujoba Alabi, Ankur Bapna, Xiang Li, Jing Zhang, Ahmed El-Kishky, Hongkun Sun, Mikel Artetxe, Abhinav Dubey, Wilker Aziz, Philipp Koehn, Sergey Edunov, Angela Fan, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics (ACL'25)*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics (ACL).

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS'21)*, Virtual. Conference on Neural Information Processing Systems (NeurIPS).

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'24)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics (ACL).

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 13541–13564, Miami, USA. Association for Computational Linguistics (ACL).

Janne Johannessen, Andre Kåsen, Kristin Hagen, Anders Nøklestad, and Joel Priestley. 2020. Comparing methods for measuring dialect similarity in Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC'20)*, pages 5343–5350, Marseille, France. European Language Resources Association (ELRA).

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics (EMNLP'23)*, pages 7226–7245, Singapore. Association for Computational Linguistics (ACL).

Daniil Larionov, Mikhail Seleznyov, Vasiliy Viskov, Alexander Panchenko, and Steffen Eger. 2024. xCOMET-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 21934–21949, Miami, USA. Association for Computational Linguistics (ACL).

Zhijian Li, Stefan Larson, and Kevin Leach. 2024. Generating hard-negative out-of-scope data with ChatGPT for intent classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, pages 7634–7646, Torino, Italia. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL).

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'22)*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics (ACL).

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, pages 1339–1352, Torino, Italia. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL).

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS'23)*, New Orleans, USA. Conference on Neural Information Processing Systems (NeurIPS).

Ananya Mukherjee, Saumitra Yadav, and Manish Shrivastava. 2025. Why should only high-resource-languages have all the fun? pivot based evaluation in low resource setting. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING'25)*, pages 4779–4788, Abu Dhabi, UAE. Association for Computational Linguistics (ACL).

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics (ACL).

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics (ACL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphia, USA. Association for Computational Linguistics (ACL).

Kyubyong Park, Yo Joong Choe, and Jiyeon Ham. 2020. Jejueo datasets for machine translation and speech synthesis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC'20)*, pages 2615–2621, Marseille, France. European Language Resources Association (ELRA).

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL'25)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics (ACL).

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics (ACL).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics (ACL).

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics (ACL).

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. Improving Jejueo-Korean translation with cross-lingual pretraining using Japanese and Korean. In *Proceedings of the 9th Workshop on Asian Translation (WAT'22)*, pages 44–50, Gyeongju, Republic of Korea. International Conference on Computational Linguistics (ICCL).

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics (ACL).

12. Language Resource References

National Information Society Agency (NIA). 2022. *Korean Dialect Data of Middle-aged and Elderly Speakers*. Lead construction institution: MTData Inc. Available at: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71517.

A. Detailed Ensemble Performance Results

Ensemble Size	Combination	Accuracy (%)
2	KcELECTRA + KoSimCSE-RoBERTa	94.92
3	KcELECTRA + KoELECTRA + KLUE-BERT	94.88
3	KcELECTRA + KoELECTRA + KLUE-RoBERTa	94.78
2	KcELECTRA + KLUE-RoBERTa	94.72
2	KcELECTRA + KLUE-BERT	94.70
2	KcELECTRA + KoELECTRA	94.68
3	KcELECTRA + KoELECTRA + KoSimCSE-RoBERTa	94.68
4	KcELECTRA + KoELECTRA + KoSimCSE-RoBERTa + KLUE-RoBERTa	94.62
4	KcELECTRA + KoELECTRA + KLUE-BERT + KoSimCSE-RoBERTa	94.56
4	KcELECTRA + KoELECTRA + KLUE-BERT + KLUE-RoBERTa	94.50
3	KcELECTRA + KLUE-BERT + KLUE-RoBERTa	94.46
5	KcELECTRA + KoELECTRA + KLUE-BERT + KoSimCSE-RoBERTa + KLUE-RoBERTa	94.40
2	KoELECTRA + KLUE-RoBERTa	94.38
2	KoELECTRA + KoSimCSE-RoBERTa	94.36
1	KcELECTRA	94.34
3	KoELECTRA + KLUE-BERT + KLUE-RoBERTa	94.22
3	KcELECTRA + KoSimCSE-RoBERTa + KLUE-RoBERTa	94.20
3	KoELECTRA + KoSimCSE-RoBERTa + KLUE-RoBERTa	94.18
4	KoELECTRA + KLUE-BERT + KoSimCSE-RoBERTa + KLUE-RoBERTa	94.16
3	KcELECTRA + KLUE-BERT + KoSimCSE-RoBERTa	94.14
4	KcELECTRA + KLUE-BERT + KoSimCSE-RoBERTa + KLUE-RoBERTa	94.12
3	KoELECTRA + KLUE-BERT + KoSimCSE-RoBERTa	94.04
1	KoELECTRA	93.98
2	KoELECTRA + KLUE-BERT	93.88
3	KLUE-BERT + KoSimCSE-RoBERTa + KLUE-RoBERTa	93.84
2	KLUE-BERT + KoSimCSE-RoBERTa	93.62
2	KoSimCSE-RoBERTa + KLUE-RoBERTa	93.56
2	KLUE-BERT + KLUE-RoBERTa	93.50
1	KLUE-RoBERTa	93.38
1	KLUE-BERT	93.20
1	KoSimCSE-RoBERTa	92.92

Table 7: Comparison of all 31 combinations constructed from five Korean pre-trained language models under identical data and hyperparameters, sorted by accuracy. The Ensemble size column indicates the number of base models (1–5).

B. Prompts

```
You are a translation assistant that translates standard Korean into {DIALECT}.
Output only the translated sentence; do not include any explanations.
```

Figure 3: System instruction used across all settings.

```
Input sentence: {SOURCE}
Translation:
```

Figure 4: Zero-shot prompt template.

```
Example:
A: {STANDARD_EXAMPLE_1}
B: {DIALECT_EXAMPLE_1}

Example:
A: {STANDARD_EXAMPLE_2}
B: {DIALECT_EXAMPLE_2}

... (total of 10 examples)

Input sentence: {SOURCE}
Translation:
```

Figure 5: In-context learning prompt template.

```
[Feedback]
- The previous output was classified as {WRONG_LABEL} instead of the target
dialect {DIALECT}.
- Please revise the translation to clearly reflect {DIALECT} features.
- Previous output : {PREV_OUTPUT}
```

Figure 6: Feedback prompt template. Used in DIA-REFINE; when an output is classified as a non-target dialect, this feedback is prepended to the base prompt (zero-shot or in-context) on the subsequent attempt to request a revision.

```
The output oscillates between {last_wrong_label} and {wrong_label}.
Please make the {dialect}-specific features more explicit.
```

Figure 7: Oscillation prompt template. Used in DIA-REFINE; when the two most recent hypotheses are classified into different non-target dialects, it is inserted into the feedback prompt to flag oscillation and help the model steer toward the target dialect.