

CebIDN: A Computational Framework for Cebuano Idiom Identification, Disambiguation, and Normalization

Adrian P. Vafflor and Robert R. Roxas

Department of Computer Science
University of the Philippines Cebu
{apvafflor1, robert.roxas}@up.edu.ph

Abstract

Idioms remain a persistent challenge for natural language understanding due to their non-compositional nature, especially in morphologically rich, low-resource languages like Cebuano. Existing work on idiom processing has explored span extraction, contrastive learning, and normalization, but typically treats these tasks independently or assumes previous steps have already been done. We propose CebIDN, a modular pipeline that unifies idiom identification, disambiguation, and normalization for Cebuano. Central to our framework is a morphology-aware canonical matcher that links extracted spans to an idiom knowledge base, passing structural tags, the idiom used, and retrieved literal hints to the downstream generative model. The system is evaluated in various tagging configurations, incorporating CRF-based sequence labeling and span-level triplet loss, along with few-shot and fine-tuned normalization strategies across both tagged and tag-free configurations. The optimal configuration (using a 3-tag BIO+CRF tagger with a tag-free normalizer) significantly outperforms a fine-tuned end-to-end LLM baseline by improving binary idiom detection (0.899 vs. 0.847 F1) and achieving superior normalization quality (88.87 vs. 86.89 BLEU; 96.40 vs. 95.04 BERTScore; 4.71 vs. 4.46 human-validated LLM evaluation scores), while running 1.7 times faster. These results demonstrate that structured task decoupling improves figurative language understanding in low-resource settings.

1 Introduction

Despite major advances in natural language processing (NLP) (Vaswani et al., 2017; Devlin et al., 2019), idiomatic expressions remain difficult to model due to their non-compositional nature (Baldwin and Kim, 2010). As a result, modern NLP systems that rely on compositional representations often interpret idioms literally (Dankers et al., 2022;

Liu et al., 2017), which can negatively affect downstream tasks such as machine translation (Fadaee et al., 2018; Salton et al., 2014), sentiment analysis (Liu et al., 2017; Biddle et al., 2020), and paraphrase generation (Zhou et al., 2022; Qiang et al., 2023).

Previous approaches to idiom processing can be broadly grouped into three directions: (i) identification and disambiguation (Muzny and Zettlemoyer, 2013; Tedeschi et al., 2022; Zeng and Bhat, 2021), (ii) contrastive representation learning for figurative-literal separation (He et al., 2024; Matheny et al., 2026), and (iii) literal paraphrasing (Zhou et al., 2022; Dai et al., 2023; Qiang et al., 2023). More recently, instruction-tuned large language models (LLMs) have been explored as a complementary approach, primarily through prompting and fine-tuning (Badran et al., 2025; Site et al., 2026). However, these approaches were largely developed in isolation or assumed that previous tasks had already been completed.

These limitations are more pronounced in low-resource languages (Pava et al., 2025). Cebuano, a widely spoken but under-resourced language in the Philippines (Eberhard et al., 2026; Tan et al., 2024; Eslit, 2025), is characterized by morphologically rich and syntactically flexible structures (Adlaon and Marcos, 2018; Fronteras et al., 2019), making it a challenging but valuable testbed for end-to-end idiom processing. In particular, studying Cebuano further enables evaluation beyond high-resource Indo-European languages, which have been the primary focus of prior work on idiom processing (Tedeschi et al., 2022; He et al., 2024).

To address these challenges, this paper presents CebIDN,¹ a two-stage pipeline for processing idioms specifically in Cebuano. Operating in a partially closed-vocabulary setting, our framework

¹Codebase architecture details, and public data/repository links are provided in Appendix A.

targets 50 Cebuano idioms covered by a curated knowledge base. The first stage performs idiom identification and disambiguation via sequence labeling to extract and classify idiomatic spans, while the second stage performs idiom normalization using a generative model to produce literal interpretations. The contributions of this work are as follows.

1. The CebIDN pipeline integrates identification and normalization, with ablations on CRF-based labeling, triplet loss (TL) and generation strategies.
2. A curated Cebuano idiom dataset covering 50 distinct idioms (9,300 sentences) with span-level BIO (Begin-Inside-Outside) annotations covering figurative, literal, and hard-negative idiom usages, along with literal interpretations.
3. Evaluation against modular LLM baselines in fine-tuned (FT) and few-shot (FS) settings.

2 Related Work

Prior work treats idiom identification and normalization as isolated tasks. To our knowledge, no previous study integrates them into a unified pipeline.

2.1 Sequence Labeling Approaches

Early systems relied on candidate-specific classification, assuming idiomatic expressions were pre-identified (Muzny and Zettlemoyer, 2013; Madabushi et al., 2021). Later work reformulated detection as sequence labeling to extract and classify spans directly from text (Tedeschi et al., 2022). While successful at disambiguation, these methods stop at identification, bypassing downstream contextual normalization (Biddle et al., 2020). For seen expressions, simple rule-based filters can perform comparably to complex neural models (Pasquer et al., 2020).

2.2 Contrastive and Representation Learning

To address this semantic gap, contrastive learning has been used to separate literal and figurative representations in embedding space. This includes triplet-based losses (He et al., 2024), span-level contrastive learning (Matheny et al., 2026), and external knowledge integration (Dai et al., 2023). While some methods combine contrastive learning with sequence tagging (Matheny et al., 2026), they are designed for high-resource settings and remain isolated from normalization models.

2.3 Idiom Normalization and End-to-End Pipelines

In parallel, idiom normalization has been studied as idiomatic sentence paraphrasing or phrasal substitution, where figurative expressions are rewritten in their literal equivalents (Liu and Hwa, 2016; Zhou et al., 2022). This task has been explored via Chinese idiom paraphrasing datasets (Qiang et al., 2023) and unsupervised multiword expression substitution techniques (Wada et al., 2023). Instruction-tuned LLMs have further improved performance in zero-shot and few-shot (FS) settings (Shu et al., 2023), while staged architectures that separate classification and generation have been shown to improve controllability under ambiguity (Badran et al., 2025). However, most existing systems, including those based on Chain-of-Thought reasoning such as ITU-NLP (Site et al., 2026), still assumed that idiomatic spans were already identified, bypassing the extraction problem and limiting robustness in fully automatic scenarios.

2.4 Low-Resource Cebuano NLP

In Cebuano, computational resources remain scarce (Alejandrino, 2016; Eslit, 2025), and broader benchmarks such as FILBENCH continue to report performance gaps in semantically complex tasks (Miranda et al., 2025). Recent work by Tan et al. (2024) introduced CebBERT, a Cebuano-specific DistilBERT model trained on a multi-domain Cebuano corpus. Their results on figurative language classification showed consistent gains over multilingual models such as mBERT, making CebBERT the closest existing line of work to Cebuano idiom processing. However, their setup was limited to sentence-level classification, and the *literal* class referred to general non-figurative sentences rather than paired literal usage of specific idioms.

Overall, prior work has shown fragmented components of idiom processing, necessitating a unified framework that works in low-resource settings.

3 CebIDN Framework

As illustrated in Figure 1, the CebIDN framework operates on a raw Cebuano sentence by executing a sequential pipeline that decouples span identification from generative normalization. Rather than processing the entire input unconditionally, the pipeline passes the raw text to Stage 1 to extract token-level idiomatic spans. When no idiomatic triggers are predicted, the sentence is classified as

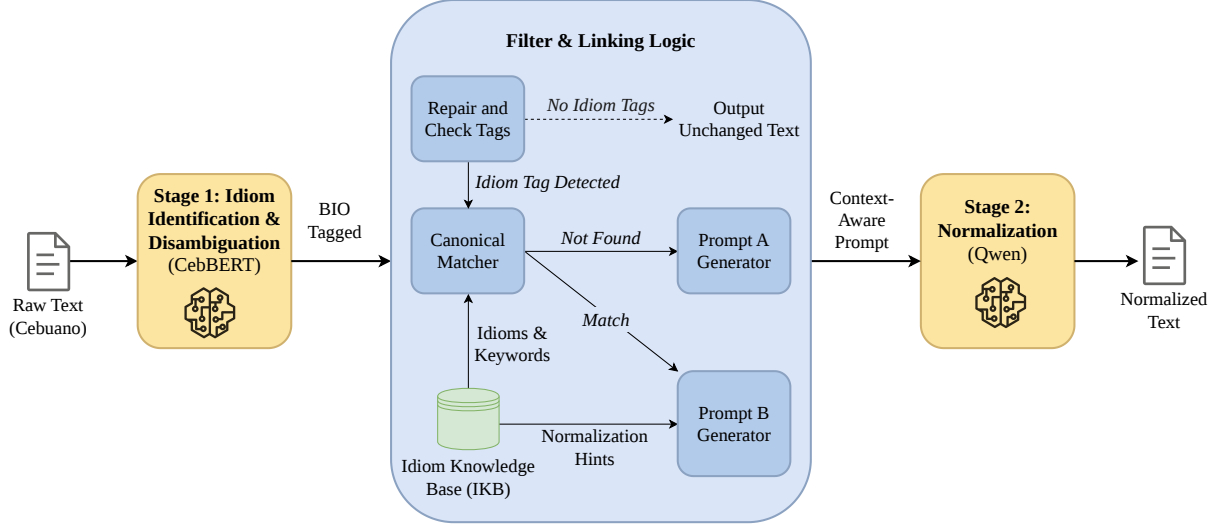


Figure 1: CebIDN inference pipeline: a two-stage process for identifying idiomatic spans and generating literal normalizations.

literal and bypassed. For sentences containing detected idioms, the predicted spans are passed to a canonical matching module to retrieve resource-grounded context from the Idiom Knowledge Base (IKB). Finally, Stage 2 utilizes the original sentence, the extracted span boundaries, and the retrieved metadata to generate the final literal paraphrased sentence.

3.1 Stage 1: Identification and Disambiguation

Stage 1 treats idiom identification and disambiguation as a joint sequence labeling task.

3.1.1 BIO Implementation

The Stage 1 identification model was fine-tuned on CebBERT (Tan et al., 2024) using a token-level BIO sequence labeling architecture following standard BERT-style fine-tuning (Devlin et al., 2019). Each token was assigned one of C BIO labels ($C = 3$ or $C = 5$ depending on the tagging scheme), and model parameters were optimized using token-level categorical cross-entropy loss.

3.1.2 Conditional Random Fields (CRF)

To enforce sequence-level consistency and valid BIO transitions, we applied a linear-chain CRF layer on top of the encoder token representations (Lafferty et al., 2001).

$$P(y|x) = \frac{\exp(\sum_{i=1}^n \psi(y_{i-1}, y_i, x))}{\sum_{y' \in \mathcal{Y}} \exp(\sum_{i=1}^n \psi(y'_{i-1}, y'_i, x))} \quad (1)$$

where x denotes the input sequence, y the label sequence, and $\mathcal{Y}$ the set of all possible label sequences. The potential function $\psi(y_{i-1}, y_i, x)$ combines encoder emission scores with learned transition scores between adjacent labels. Training maximizes the conditional log-likelihood of the gold sequence. During inference, standard unconstrained Viterbi decoding is applied. Any remaining illegal BIO transitions, such as an I- tag without a preceding B- tag, are subsequently resolved in the downstream filter and linking layer (Section 3.2).

3.1.3 Triplet Loss (TL) and Contrastive Learning

To improve figurative-literal disambiguation, span-level contrastive learning was applied over pooled representations of idiomatic spans. The representation of a span S was computed via mean pooling:

$$z_{span} = \frac{1}{|S|} \sum_{t \in S} h_t \quad (2)$$

where S denotes the idiomatic span (excluding surrounding context tokens) and h_t is the hidden representation of token t .

Triples (a, p, n) consisted of an anchor span a , a positive span p sharing the same interpretation type (figurative or literal), and a negative span n with the opposite interpretation type.

Unlike standard triplet construction, positives and negatives were sampled strictly within the same idiom identity to prevent the model from exploiting inter-idiom lexical differences as shortcut cues. While this constraint focuses representation learn-

ing on the target corpus and does not explicitly optimize for generalization to unseen idioms, we prioritize this configuration to maximize the model’s disambiguation performance. Evaluating contrastive representation generalization on unseen idioms is left for future work.

Following the standard triplet margin loss formulation (Schroff et al., 2015), the contrastive objective was formulated as:

$$L_{TL} = \max(d(a, p) - d(a, n) + m, 0) \quad (3)$$

where $d(\cdot)$ denotes the Euclidean distance and m is a margin hyperparameter fixed to 1.0. The final multitask objective combined the sequence labeling and contrastive losses:

$$L_{total} = L_{BIO/CRF} + \alpha \cdot L_{TL} \quad (4)$$

where α was a weighting coefficient tuned on the validation set.

3.2 Filter and Linking Logic

The transition between identification and normalization was managed by a filtering and linking layer. Sentences containing only literal spans or without predicted idiomatic labels were discarded. Before canonical matching, invalid BIO sequences were parsed and resolved into valid span boundaries during entity extraction using *sequeval* (Nakayama, 2018)

3.2.1 Canonical Matcher and Knowledge Retrieval

Predicted idiomatic spans were linked to a curated Idiom Knowledge Base (IKB) via a morphology-aware canonical matching algorithm. The IKB contains a list of idioms, their matching keywords, and normalization hints consisting of candidate literal replacements. Standard exact-string matching was insufficient for Cebuano, where lexical roots can surface in multiple inflected forms through affixation (e.g., *nag-*, *gi-*, *-an*, *-on*). The matcher therefore employed root-inclusion substring matching: a keyword k matched a span token t when k appeared as a case-insensitive substring of t , allowing inflected forms to be recognized without exhaustive enumeration of morphological variants.

As illustrated in Table 1, keywords were organized into OR-based groups capturing lexical, morphological, and common spelling variants. Curated keyword sets retained only informative lexical anchors, preserving the minimal cues necessary for

Idiom (+ EN Meanings)	Keyword Groups
<i>abli ang gaisano</i> (Lit: Gaisano mall is open; Fig: fly/zipper is open)	<i>abli/abri, gaisano</i>
<i>baga’g nawong</i> (Lit: thick face; Fig: shameless)	<i>baga, nawng/nawong</i>
<i>hinaw</i> (Lit: to wash hands; Fig: to absolve oneself)	<i>hinaw/hunaw</i>

Table 1: Example keyword groups used for morphology-aware canonical matching. Slash-separated forms indicate OR logic over lexical, morphological, or spelling variants. Keyword groups are combined using AND logic.

idiom identification, since some expressions remain idiomatic even when non-essential tokens are omitted in reduced surface forms. Frequent grammatical particles and case markers (e.g., *ang*, *sa*, *og*, *nga*, *pa*) were excluded from the keyword set, restricting matching to lexical content words. Prior to matching, lightweight normalization removed case distinctions and apostrophes commonly used in Cebuano contracted forms. All required keyword groups had to match within a span (AND logic) to ensure the idiomatic expression was present.

Successful matches retrieved the canonical idiom form and normalization hints used during normalization. In pilot evaluations on gold-standard annotated spans, the matcher achieved 100% coverage across all 50 idioms. While this perfect coverage was guaranteed by design because the IKB keyword groups were curated in tandem with the dataset sentences to ensure successful canonical retrieval, applying rule-based matching directly to raw sentences is highly unreliable in real-world scenarios due to contextual ambiguity (literal and hard-negative usages) and morphological noise. Placing the sequence tagger first isolates the candidate figurative span, reducing the search space. If a real-world span fails the keyword rules due to unseen variations, dialectal shifts, typos, or other uncovered word and spelling variations, the pipeline bypasses the dictionary and routes the sentence to the normalization prompt format without hints.

3.3 Stage 2: Normalization Phase

The second stage converted identified idiomatic spans into grammatically integrated literal equivalents using a fine-tuned Qwen-2.5-7B generative model (Yang et al., 2024). Unlike end-to-end idiom normalization, where the model must jointly

detect idiomatic expressions and infer appropriate literal substitutions from internal knowledge, Stage 2 operated on a simplified conditional generation task focused primarily on literal span replacement and sentence-level grammatical integration.

The prompt consisted of a task instruction, the original sentence (either wrapped using `<idiom>... </idiom>` tags or provided tag-free), and optional dictionary-derived metadata. When a canonical match was available, the prompt was augmented with the canonical idiom form and normalization hints retrieved from the IKB (Prompt B in Figure 1). Otherwise, the model received only the tagged or raw sentence (Prompt A), allowing the fine-tuned normalizer to rely on learned semantic representations for unresolved or unseen variants. Full prompt templates are provided in Appendix H.

4 Dataset Construction

To support the end-to-end pipeline, a new Cebuano idiom corpus was constructed to enable integrated processing.

4.1 Corpus Design

The corpus was constructed from 50 idioms spanning three contextual sentence categories: (1) *figurative* (paired with a *normalized literal form*), (2) *literal* (non-figurative usage of the same expression), and (3) *hard negative* (HN; non-idiomatic constructions with partial lexical overlap). Idioms were divided into three types based on the identified categories:

- **Complete:** Figurative, literal, and hard negative forms are available, e.g., *abli ang gaisano* (‘fly/zipper is open’; literal: ‘Gaisano mall is open’; HN: *abli ang bag*, ‘bag is open’).
- **No-Literal:** No meaningful literal interpretation is available, e.g., *anak sa liki nga kawayan* (‘child born out of wedlock’; HN: *liki nga kawayan*, ‘cracked bamboo’).
- **No-HN:** Single-word idioms with literal and figurative readings but no separable hard negative, e.g., *hinaw* (‘to absolve oneself’; literal: ‘to wash hands’).

The corpus also included two pairs of semantically related idioms sharing similar figurative meanings, designed to discourage lexical memorization and require contextual disambiguation (for example, *baga’g lips* and *baga’g nawong*, which literally mean ‘thick lips’ and ‘thick face’ and both figuratively denote ‘shameless’; and *anak sa gawas* and

anak sa liki nga kawayan, which literally mean ‘child outside’ and ‘child of a cracked bamboo’ and both figuratively denote an ‘illegitimate child’). Each idiom was further annotated with grouped lexical keywords and normalization hints for use in canonical matching (Section 3.2.1) and retrieval-augmented prompting (Section 3.3).

4.2 Data Collection and Validation

The idiom entries and their meanings were collected from a Kaggle Cebuano idiom list (Bestal, 2024), originally derived from Wiktionary (Wiktionary contributors, 2016), and subsequently filtered and validated by a native speaker. For idioms with multiple figurative meanings, the most commonly attested meaning was selected. Sentences were constructed through a combination of manual authoring by a native Cebuano speaker (~50%) and AI-assisted generation using ChatGPT and Gemini variants (OpenAI, 2022; Google, 2023) (~50%), with all AI-generated outputs manually reviewed, corrected, and revised prior to inclusion. Instances comprise 1-5 short sentences with sufficient context for disambiguation.

All human evaluations in this work were conducted by native-level Cebuano speakers (see Appendix B for recruitment and demographic details). For the dataset quality check, a randomly sampled subset of figurative-normalization pairs (two per idiom) was independently rated by three evaluators on 5-point Adequacy and Fluency scales. Ratings were high (4.96 ± 0.19 ; 4.93 ± 0.29) with strong pairwise agreement (92.67%; 90.33%). Fleiss’ κ was lower (-0.038 ; 0.104), consistent with reported sensitivity to ceiling effects in skewed annotation distributions (Feinstein and Cicchetti, 1990).

To further verify the quality and category consistency of the dataset, three native-level Cebuano reviewers independently classified a randomly selected validation sample of 136 sentences representing all three contextual categories (figurative, literal, and hard negative). Reviewer classifications demonstrated near-perfect alignment with the gold standard labels, yielding an average accuracy of 99.51% (with two reviewers achieving 100% accuracy and the third achieving 98.53%). Inter-rater agreement was extremely high, with an average pairwise percentage agreement of 99.02%, an average pairwise Cohen’s κ of 0.985, and a multi-rater Fleiss’ κ of 0.985. Absolute consensus across all three reviewers was achieved on 98.53% (134/136) of the sample, confirming the distinct semantic

Idiom Type	Idioms	Train	Val	Test	Total
Complete	36	5,760	720	720	7,200
No-Literal	10	1,200	150	150	1,500
No-HN	4	480	60	60	600
Total	50	7,440	930	930	9,300

Table 2: Dataset composition and manually curated splits across the three corpus categories.

boundaries between literal, figurative, and hard-negative usages in the dataset.

4.3 Dataset Counts and Splits

Dataset splits were manually curated to preserve representative distributions of usage types and difficulty levels across partitions. For each idiom, 50 instances were generated per usage category using an 80/10/10 split (40 train, 5 validation, 5 test), resulting in 200 instances per Complete idiom and 150 per No-Literal or No-HN idiom, for a total of 9,300 instances. Table 2 summarizes the dataset.

5 Experimental Setup

For Stage 1, we evaluate four identification architectures: BIO, BIO+CRF, and their triplet loss (TL) counterparts, BIO+TL and BIO+CRF+TL, under both a 5-label scheme (B/I-IDIOM, B/I-LITERAL, O) and a collapsed 3-label scheme mapping literal spans to 0. The TL objective was applied only in the 5-label setting, as the 3-label formulation collapses literal spans into the background class. The contrastive weight α was tuned on the validation split, yielding an optimal value of $\alpha = 0.05$; detailed sweep results are provided in Appendix C.3.

For Stage 2, Qwen-2.5-7B was evaluated under eight normalization configurations: zero-shot prompting (ZS), 3-shot prompting (FS), 3-shot prompting with hints (FS+hints), 3-shot prompting with hints and tags (FS+hints+tags), fine-tuning (FT), fine-tuning with hints (FT+hints), fine-tuning with tags (FT+tags), jointly trained with FT+hints+tags), and fine-tuning with hints and tags (FT+hints+tags).

These configurations were compared against monolithic end-to-end Qwen-2.5-7B baselines in zero-shot (E2E ZS), 7-shot prompting (E2E FS), and fine-tuned (E2E FT) settings, as well as retrieval-augmented generation baselines with oracle hints in both 7-shot prompting (RAG FS) and fine-tuned (RAG FT) settings. In the RAG configurations, the prompt included the candidate idiom and its corresponding normalization hints, leaving it to the model to contextually determine whether to apply the literal normalization or bypass it. All

Stage 1 sequence labeling, Stage 2 normalization, and end-to-end baseline experiments were conducted on Kaggle using dual T4 GPUs across three random seeds (42, 123, and 456), with results reported as mean $\pm$ standard deviation. Complete hyperparameter details are provided in Appendix C.

5.1 Evaluation Metrics

Stage 1 idiom identification was evaluated using three structural and span-level metrics. Trigger F1 measures the exact-match extraction of idiomatic spans, computed using seqeval on idiom-only tags. For 5-tag configurations, literal tags are collapsed into the background category prior to F1 calculation. Invalid BIO sequences (Inv. BIO) count the absolute number of syntactically illegal transitions in the raw model predictions, such as an I- tag appearing without a preceding B- tag, prior to any heuristic correction. Partial matches (Partial) quantify boundary-mismatched idiom detections, counting cases where a predicted span overlaps with a ground-truth idiomatic span but fails to align with its exact boundaries.

Stage 2 normalization quality was evaluated using three complementary metrics. BLEU computes n -gram overlap with the reference, utilizing SacreBLEU. BERTScore measures semantic similarity using the contextual token embeddings of the raw, pretrained CebBERT encoder. The composite score represents the average of three quality dimensions rated on a 1–5 scale by an LLM-as-a-Judge using GPT-4o mini (OpenAI, 2024). These dimensions comprise Semantic Fidelity for meaning preservation, Span Accuracy for the precise replacement of only the idiomatic span, and Fluency for grammatical naturalness.

To validate the LLM-as-a-Judge framework, three native Cebuano speakers independently rated 100 randomly sampled candidates from the seed 42 validation split, excluding simple copies and exact match items. Human and LLM agreement, averaged across the three raters, yielded strong Spearman correlations, specifically $\rho = 0.791$ for Semantic Fidelity, $\rho = 0.690$ for Span Accuracy, $\rho = 0.498$ for Fluency, and an overall composite correlation of $\rho = 0.702$ (all $p < 0.0001$). Weighted Cohen’s κ scores were 0.797, 0.509, and 0.420, respectively, confirming substantial alignment on semantic fidelity and moderate alignment on span accuracy and fluency. To eliminate evaluator variance and prevent potential LLM hallucinations, exact matches to the expected gold nor-

Tags	Config	Trigger F1	Inv. BIO	Partial
3-tag	BIO	0.834 $\pm$ 0.017	16.0 $\pm$ 7.8	15.0 $\pm$ 6.6
3-tag	BIO+CRF	0.876 $\pm$ 0.011	5.7 $\pm$ 2.9	5.3 $\pm$ 2.1
5-tag	BIO	0.852 $\pm$ 0.003	27.7 $\pm$ 5.5	7.7 $\pm$ 1.5
5-tag	BIO+CRF	0.877 $\pm$ 0.015	9.7 $\pm$ 6.4	2.0 $\pm$ 1.0
5-tag	BIO+TL	0.861 $\pm$ 0.012	32.3 $\pm$ 2.5	7.3 $\pm$ 2.1
5-tag	BIO+CRF+TL	0.874 $\pm$ 0.009	10.3 $\pm$ 2.9	3.0 $\pm$ 2.0

Table 3: Stage 1 (CebBERT) idiom identification performance. TL denotes Triplet Loss. Best results within each tagging scheme are highlighted in bold.

malization were automatically pre-assigned perfect scores across all dimensions. Conversely, outputs that merely copied the input sentence verbatim by-passed LLM evaluation and were pre-assigned a score of 1 for semantic fidelity, 1 for span accuracy, and 5 for fluency, while still being evaluated via standard BLEU and BERTScore. Furthermore, during multi-candidate queries, the presentation order of candidates was randomized prior to prompting to eliminate positional bias.

Finally, the pipeline was evaluated under four end-to-end metrics. Binary F1 (Bin. F1) measures sequence-level figurative classification regardless of span boundaries. To isolate Stage 2 quality, BLEU, BERTScore, and the Composite score were computed exclusively on Stage 1 True Positive (TP) cases. Average Quality is a holistic per-item metric (scaled 0–100) assigning full credit to correct decisions and zero to errors. Pipeline Speed measures average execution latency in milliseconds per sentence. See Appendix G for detailed calculations.

6 Results and Discussions

Experimental results are provided for Stage 1 idiom identification and Stage 2 normalization, along with ablation studies and comparisons against end-to-end LLM baselines.

6.1 Stage 1: Idiom Identification

Table 3 summarizes Stage 1 performance. We present the full classification breakdown including precision and recall scores for each tagging scheme in Appendix D. Trigger F1 strongly inversely correlated with partial matches ($r = -0.95$; Appendix Figure D.1), indicating that performance differences were driven mainly by boundary precision rather than idiom sensitivity. Since partial spans are still forwarded to normalization, weaker Trigger F1 does not necessarily imply weaker downstream performance.

Effect of CRF CRF decoding was the most effective architectural enhancement across both tagging schemes, mitigating transition ambiguity through global sequence constraints. Under the 3-tag setting, BIO+CRF improved Trigger F1 by 4.2 points while reducing invalid sequences and partial-span errors by roughly two-thirds. In 5-tag setting, CRF yielded a 2.5-point Trigger F1 gain alongside substantial reductions in transition violations, demonstrating the value of sequence-level constraints on more complex label vocabularies. While CRF decoding significantly reduces transition violations, it does not eliminate them entirely due to the unconstrained Viterbi decoding applied at inference (Section 3.1.2), where strong token-level emissions can occasionally override learned transition penalties.

Effect of Triplet Loss While triplet loss maximizes sensitivity to figurative spans, yielding the highest recall of 0.911 (Appendix D.1), semantic supervision alone does not guarantee precise span boundaries. Consequently, BIO+TL produced the highest rate of invalid transitions alongside high boundary mismatches. Integrating CRF decoding (BIO+CRF+TL) effectively resolves these structural issues, although Trigger F1 remains slightly below the standard BIO+CRF baseline, highlighting a direct trade-off between recall and boundary precision.

3-tag vs. 5-tag trade-off Explicit literal-figurative separation (5-tag) improves span detection by 1.8 Trigger F1 points under unconstrained decoding, but this advantage vanishes when constrained by CRF decoding. The two schemes still present distinct structural trade-offs: the 5-tag BIO+CRF configuration yields higher boundary precision with fewer partial-span matches, whereas the 3-tag counterpart favors decoding stability with fewer invalid transitions.

6.2 Stage 2: Normalization Quality

Stage 2 normalization is evaluated under both few-shot and fine-tuned settings to understand how retrieved lexical metadata and structural span conditioning influence literal rewriting quality.

Normalization in perfect spans Table 4 shows that under untuned settings, the zero-shot and few-shot baselines are highly unreliable. We present the complete multidimensional normalization evaluation results including ROUGE-L metrics in Ap-

Config	BLEU	BERTScore	Composite
ZS	48.62 $\pm$ 0.00	76.94 $\pm$ 0.00	1.98 $\pm$ 0.00
FS	53.83 $\pm$ 0.22	79.26 $\pm$ 0.06	2.00 $\pm$ 0.01
FS + hints	73.16 $\pm$ 0.63	90.38 $\pm$ 0.15	3.66 $\pm$ 0.06
FS + hints + tags	76.53 $\pm$ 0.64	91.88 $\pm$ 0.20	4.01 $\pm$ 0.03
FT	86.10 $\pm$ 0.63	94.57 $\pm$ 0.49	4.39 $\pm$ 0.07
FT + hints	88.51 $\pm$ 0.25	96.37 $\pm$ 0.03	4.74 $\pm$ 0.01
FT + tags	86.00 $\pm$ 0.84	94.81 $\pm$ 0.42	4.42 $\pm$ 0.01
FT + hints + tags	88.51 $\pm$ 0.59	96.18 $\pm$ 0.06	4.69 $\pm$ 0.02

Table 4: Stage 2 normalization performance across configurations. All configurations utilize Qwen-2.5-7B as the generative backbone. Few-shot settings use 3 in-context examples. Best results per metric are highlighted in bold.

pendix E. Generative normalization either copies the input sentence verbatim or produces nonsensical translations through syntax hallucinations, unknown word insertions, and semantic leakage from few-shot exemplars (detailed in Appendix J; Table J.1). This indicates that the base model cannot independently resolve figurative Cebuano idioms. Tagged variants of few-shot and zero-shot configurations are therefore omitted, as their untuned counterparts already demonstrate insufficient pre-trained knowledge. However, adding dictionary normalization hints improves the few-shot Composite score by 1.66 points (19.33 BLEU points), which is further enhanced by 0.35 Composite points (3.37 BLEU points) when the idiomatic span is explicitly marked using boundary tags, highlighting the relevance of the IKB.

The importance of the IKB also applies to fine-tuned configurations. Even without dictionary hints, plain fine-tuning outperforms few-shot baselines. However, the influence of explicit boundary tags becomes less pronounced under fine-tuning, leaving dictionary hints as the most valuable addition. This is primarily because tagged configurations are fragile, occasionally leaking boundary tags into the generated output even when input boundaries are perfect (Appendix K). Because the presence of markup tokens in the final translation is heavily penalized by the LLM-as-a-Judge prompt (Appendix I), tagged configurations suffer from quality penalties. A tag-free setup makes the normalizer resistant to such formatting errors, allowing it to rely on its learned representations without having to explicitly output or strip tags.

Normalized partial-match quality Across all experimental runs, Stage 1 produced 96 unique partial-span matches, with 17 instances successfully matched to canonical entries in the IKB and

Config	BLEU	BERTScore	Composite
<i>With Canonical Match</i>			
ZS	58.96 $\pm$ 0.00	81.41 $\pm$ 0.00	2.12 $\pm$ 0.00
FS	56.42 $\pm$ 0.40	80.54 $\pm$ 0.18	2.07 $\pm$ 0.02
FS + hints	76.86 $\pm$ 0.41	92.12 $\pm$ 0.20	3.89 $\pm$ 0.18
FS + hints + tags	77.88 $\pm$ 0.81	91.88 $\pm$ 0.75	3.82 $\pm$ 0.15
FT	85.98 $\pm$ 2.05	94.10 $\pm$ 1.04	4.04 $\pm$ 0.35
FT + hints	90.03 $\pm$ 1.50	96.58 $\pm$ 0.33	4.73 $\pm$ 0.07
FT + tags	70.29 $\pm$ 7.57	86.01 $\pm$ 5.22	2.94 $\pm$ 0.34
FT + hints + tags	77.08 $\pm$ 2.77	90.51 $\pm$ 2.44	3.69 $\pm$ 0.18
<i>Without Canonical Match</i>			
ZS	45.03 $\pm$ 0.00	75.29 $\pm$ 0.00	1.92 $\pm$ 0.00
FS	50.94 $\pm$ 1.07	78.11 $\pm$ 0.63	1.97 $\pm$ 0.07
FT	83.50 $\pm$ 2.28	92.54 $\pm$ 1.31	3.98 $\pm$ 0.16
FT + tags	67.17 $\pm$ 2.53	86.11 $\pm$ 0.74	2.95 $\pm$ 0.02

Table 5: Normalization quality on partial-match Stage 1 results across FS (3-shot) and FT configurations, utilizing Qwen-2.5-7B as the generative backbone. Subsets are partitioned based on whether the canonical idiom was recoverable from the predicted span. Best results within each metric per group are highlighted in bold.

79 left unmatched. Given these small sample sizes, the analysis here remains tentative and may not generalize to larger distributions.

Table 5 reports the normalization quality on partial-match cases. Across both subsets, FT + hints and FT are the best configurations for groups with and without canonical matches, respectively, primarily due to their resilience against boundary tag formatting errors (Appendix K).

In the few-shot (FS) setting, adding hints and boundary tags substantially improves performance, allowing the FS baseline to surpass the fine-tuned configuration with tags (FT + tags). This suggests that few-shot prompting benefits from explicit structural guidance when the partial span is sufficient to retrieve the canonical idiom. Conversely, because the fine-tuned normalizer is trained on perfect gold boundaries, the presence of imperfect boundary tags on partial matches severely degrades the quality.

6.3 CebIDN vs. E2E LLMs

Table 6 compares pipeline configurations with end-to-end baselines. All pipeline variants substantially outperformed E2E baselines (even with the tagged normalizer setup in Appendix F), with the best configuration (3-tag BIO+CRF) improving over the fine-tuned E2E model by over 4 Avg. Quality points. Appendix L presents a detailed error analysis across idiom types, morphological variations, and pipeline propagation modes. Although the oracle RAG baseline slightly outperforms CebIDN (3-

Config	Bin. F1	BLEU	BERTScore	Composite	Avg. Quality	Speed (ms/sent)
3-tag BIO	0.894 $\pm$ 0.001	88.85 $\pm$ 0.57	96.33 $\pm$ 0.20	4.70 $\pm$ 0.03	89.30 $\pm$ 0.38	535 $\pm$ 90
3-tag BIO+CRF	0.899 $\pm$ 0.009	88.87 $\pm$ 0.51	96.40 $\pm$ 0.12	4.71 $\pm$ 0.03	89.82 $\pm$ 0.72	530 $\pm$ 89
5-tag BIO	0.886 $\pm$ 0.005	89.12 $\pm$ 0.68	96.47 $\pm$ 0.17	4.73 $\pm$ 0.03	89.05 $\pm$ 0.50	524 $\pm$ 96
5-tag BIO+CRF	0.886 $\pm$ 0.013	89.13 $\pm$ 0.25	96.53 $\pm$ 0.06	4.74 $\pm$ 0.02	88.97 $\pm$ 1.31	531 $\pm$ 64
5-tag BIO+TL	0.893 $\pm$ 0.016	88.87 $\pm$ 0.53	96.39 $\pm$ 0.08	4.72 $\pm$ 0.02	89.12 $\pm$ 1.55	551 $\pm$ 88
5-tag BIO+CRF+TL	0.890 $\pm$ 0.013	89.23 $\pm$ 0.25	96.49 $\pm$ 0.02	4.73 $\pm$ 0.01	89.47 $\pm$ 1.24	513 $\pm$ 93
E2E ZS	0.459 $\pm$ 0.000	53.02 $\pm$ 0.00	79.44 $\pm$ 0.00	2.13 $\pm$ 0.00	33.28 $\pm$ 0.00	551 $\pm$ 1
E2E FS	0.428 $\pm$ 0.022	54.35 $\pm$ 0.68	79.51 $\pm$ 0.52	2.08 $\pm$ 0.04	53.02 $\pm$ 1.46	1166 $\pm$ 206
E2E FT	0.847 $\pm$ 0.016	86.89 $\pm$ 0.89	95.04 $\pm$ 0.40	4.46 $\pm$ 0.04	85.48 $\pm$ 1.21	924 $\pm$ 214
RAG (Oracle) FS	0.544 $\pm$ 0.003	74.18 $\pm$ 0.29	90.50 $\pm$ 0.16	3.85 $\pm$ 0.05	34.89 $\pm$ 0.83	1666 $\pm$ 197
RAG (Oracle) FT	0.869 $\pm$ 0.022	89.21 $\pm$ 0.66	96.39 $\pm$ 0.21	4.72 $\pm$ 0.02	87.93 $\pm$ 1.35	929 $\pm$ 118

Table 6: End-to-end system comparison. TL denotes Triplet Loss. All generative configurations utilize Qwen-2.5-7B as the backbone. Few-shot (FS) configurations use 7 in-context examples. CebIDN configurations utilize CebBERT for Stage 1 sequence labeling and Qwen-2.5-7B with adaptive routing (Fine-Tuned with hints and Fine-Tuned only) for Stage 2 normalization. Overall best results per metric are highlighted in bold.

tag BIO+CRF) on the quality of successfully normalized sentences due to prepending gold definitions as hints, our decoupled architecture achieves a higher overall Binary F1, superior Average Quality, and the latency improvements. This shows that monolithic prompt-inserted definitions are inferior to selective task decomposition.

Decoupled configurations also present a quality-speed trade-off depending on the routing scheme. Tag-free normalizers (Table 6) yield superior translation quality across all configurations. However, routing with explicit boundary tags (Appendix F.1) is approximately 60 ms/sent faster inference, as tag markup constrains the generative editing domain. We present qualitative success case studies across both tagged and tag-free routing configurations in Appendix M (Table M.1).

Computationally, the pipeline achieves 1.7 times faster inference under the optimal configuration (3-tag BIO+CRF) and up to 2.0 times faster under the fastest configuration (5-tag BIO+CRF+TL under tagged routing) compared to the fine-tuned end-to-end model. This gain stems from selective inference, where the generative normalizer is only triggered for sentences containing predicted figurative spans. We detail the mathematical derivation and variables used to compute the overall pipeline latency in Appendix G.

7 Conclusion and Future Work

In this work, we introduced CebIDN, a modular pipeline for end-to-end Cebuano idiom identification, disambiguation, and normalization, supported by a newly constructed corpus of 9,300 sentences. Our results demonstrate that even large multilingual LLMs exhibit limited native capability in low-

resource idiomatic understanding, and that task decomposition using smaller, language-specific models is more effective than raw model scaling. By decoupling token-level span identification from generative rewriting, the optimal CebIDN configuration (3-tag BIO+ CRF combined with a tag-free FT+hints normalizer) achieves superior normalization quality while running 1.7 times faster than fine-tuned, single-prompt end-to-end models. Within the pipeline, integrating CRF decoding is crucial for boundary precision, and contrastive triplet loss improves figurative recall at the cost of span precision. For the generative normalizer, fine-tuning without explicit boundary markers (tag-free routing) combined with retrieved dictionary hints from an Idiom Knowledge Base yields the highest semantic accuracy and robustness to formatting errors. Future work will explore improving matching robustness to non-standard dialectal spelling, fine-tuning the normalizer on predicted noisy spans to mitigate error propagation, evaluating alternative encoder and generative model architectures, testing generalization to unseen idioms, and extending this decoupled framework to other low-resource language families.

Limitations

While CebIDN provides a robust and computationally efficient framework for Cebuano idiom processing, several limitations should be noted.

Dataset Constraints The curated Cebuano idiom dataset covers 50 idioms (9,300 sentences), and the evaluation does not determine the minimum data count or annotations per idiom required to achieve optimal system performance. Further-

more, each figurative instance has only a single gold-standard normalized counterpart, which was decided by only a single native Cebuano speaker, limiting paraphrase diversity. Additionally, each sentence in the dataset contains at most a single candidate idiomatic phrase with a single semantic usage (either purely literal or purely figurative), excluding sentences with multiple co-occurring idioms or mixed figurative and literal usages.

Rule-Based Keyword Matching The canonical matching module relies on a morphology-aware substring rule matching strategy. While this achieved 100% coverage on the test set, rule-based keyword groups are inherently static and may struggle with highly non-standard dialectal spelling variations or complex morphological shifts not covered by these rules.

Boundary Error Sensitivity The pipeline is sensitive to Stage 1 boundary errors. In tagged normalizer configurations, incorrect XML span tags propagate to Stage 2, leading to tag leakage and syntactic corruption. While our optimal tag-free configuration is immune to tag leakage, boundary errors can still degrade its normalization quality by preventing the canonical matcher from retrieving dictionary hints for the generative model.

Generalization to Unseen Idioms Because our sequence tagger (Stage 1) and Idiom Knowledge Base (Stage 2 linking) were trained and curated on a closed vocabulary of 50 idioms, the system’s ability to generalize to out-of-vocabulary (unseen) idioms is constrained. Furthermore, the span-level contrastive triplet loss was trained by sampling positive and negative instances strictly within the same idiom identity to prevent lexical shortcut learning, which does not optimize the encoder representations for unseen expressions. Consequently, for unseen idioms, Stage 1 may fail to detect the figurative span due to lack of exposure during training, and the canonical matcher will fail to retrieve normalization hints since the IKB is static. However, the pipeline’s architectural fallback path mitigates this: if an unseen span is successfully detected, it is routed to the normalizer without hints (Prompt A). While the generative model can theoretically rely on its pre-trained open-vocabulary capabilities to normalize the sentence, we expect performance to degrade compared to the 50 target idioms. A systematic evaluation of generalization to unseen idioms remains an objective for future work.

Model Scope The system was evaluated using a single encoder backbone (CebBERT) and a single instruction-tuned generative model (Qwen-2.5-7B-Instruct). While results demonstrate strong performance under this setup, the effectiveness of CebIDN across alternative pretrained encoders and larger or smaller generative models remains untested.

Hyperparameter Tuning The contrastive weight hyperparameter α sweep was evaluated using only a single training seed (Seed 42), which may limit its generalizability across different model initializations. We detail these sweep results in Appendix C.3.

Ethical Considerations

This work complies with the ACL Code of Ethics. We address key ethical considerations regarding our dataset construction, human participation, offensive content, and model usage below.

Human Participation and Consent Our dataset validation and system evaluations involved native Cebuano speakers acting as annotators. Participation was entirely voluntary, and no coercion was applied. All participants were fully briefed on the nature and goals of the study and provided informed consent prior to evaluating the sentences. No personally identifiable information was collected during this process.

Offensive and Sensitive Language Because idiomatic expressions in natural language, as well as the contexts in which they are used, can sometimes convey sarcasm, informal registers, or sensitive/offensive concepts (such as swear words, insults, or informal curses), some entries and exemplifying sentences in our dataset contain colloquial Cebuano expressions. These reflect real-world Cebuano usage and are necessary to accurately model figurative language. Any unnecessary or out-of-context offensive usages were filtered out during dataset construction and manual validation. These colloquialisms were not deemed to pose any psychological risk or harm to the evaluators, and their inclusion in the dataset is strictly for computational linguistic research and does not reflect the views or opinions of the dataset creator or the authors.

Use of AI-Generated Data and Models Approximately 50% of the sentences in our corpus were initially drafted using commercial language models

(ChatGPT and Gemini). A native Cebuano speaker manually reviewed and revised all generated text to ensure linguistic naturalness and mitigate potential model-introduced artifacts or biases. Fine-tuning and inference were conducted using publicly accessible academic compute resources (Kaggle dual T4 GPUs), the source Wiktionary/Kaggle idiom lists (publicly available under CC BY-SA), and open-weight models (Qwen-2.5-7B and CebBERT), all of which were used in accordance with their respective licenses and terms of use.

Acknowledgements

We thank Ma. Ares Vafloor, Neil Lim, Edeliza Alindao, Rona Mae Andales, Alejandro Caballes IV, Norvelyn Khym Socorin, and those who chose to remain anonymous for their various contributions, including dataset evaluation, resource sharing, and constructive feedback.

References

- Kristine Mae M. Adlaon and Nelson Marcos. 2018. [Neural Machine Translation for Cebuano to Tagalog with Subword Unit Translation](#). In *2018 International Conference on Asian Language Processing (IALP)*, pages 328–333, Bandung, Indonesia. IEEE.
- Kimberly Jhie Alejandrino. 2016. [Pagtuki sa Binalaybay: A Stylistic Analysis of Sebuano Figurative Language in Iligan National Writers Workshop's Sebuano Poems](#). *Langkit : Journal of Social Sciences and Humanities*, 6:68–108.
- Mohamed Badran, Youssef Nawar, and Nagwa El Makky. 2025. [AlexUNLP-NB at SemEval-2025 Task 1: A Pipeline for Idiom Disambiguation and Visual Representation](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 546–550, Vienna, Austria. Association for Computational Linguistics.
- Timothy Baldwin and Su Nam Kim. 2010. [Multiword expressions](#). In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2nd edition, pages 267–292. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA.
- John William Bestal. 2024. Bisaya idiom english meanings dataset. <https://www.kaggle.com/datasets/johnwilliambestal/bisaya-idiom-english-meaning>. Kaggle dataset, MIT license, accessed: 2026-05-18.
- Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. [Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter](#). In *Proceedings of The Web Conference 2020*, pages 1217–1227. Association for Computing Machinery (ACM).
- Yu Dai, Yuqiao Liu, Lei Yang, and Yufan Fu. 2023. [An Idiom Reading Comprehension Model Based on Multi-Granularity Reasoning and Paraphrase Expansion](#). *Applied Sciences*, 13(9):5777.
- Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. [Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2026. [Ethnologue: Languages of the world. Cebuano](#). Online version; accessed 15-May-2026.
- Edgar Eslit. 2025. [Bisaya 2.0: Revitalizing the Heart of a Language with AI and Cultural Innovation](#).
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. [Examining the Tip of the Iceberg: A Data Set for Idiom Translation](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. [High agreement but low Kappa: I. the problems of two paradoxes](#). *Journal of Clinical Epidemiology*, 43(6):543–549.
- Michaela Fronteras, Mary Ann Malimas, and Charis Pielago. 2019. [Cebuano Particles and their Functions in Selected Radio Segments](#). *Central Mindanao University Journal of Science*, 23(1).
- Google. 2023. [Gemini](#). Accessed: 2026-05-24.
- Wei He, Scarton Carolina, Marco Idiart, and Aline Villavicencio. 2024. [Enhancing Idiomatic Representation in Multiple Languages via an Adaptive Contrastive Triplet Loss](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. [Conditional random fields: Probabilistic models for segmenting and labeling sequence data](#). In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

- Changsheng Liu and Rebecca Hwa. 2016. [Phrasal substitution of idiomatic expressions](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373.
- Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. [Idiom-Aware Compositional Distributed Semantics](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.
- Harish Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. [ASTitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Blake Matheny, Phuong Minh Nguyen, and Minh Le Nguyen. 2026. [Span Modeling for Idiomaticity and Figurative Language Detection with Span Contrastive Loss](#). *arXiv preprint*. ArXiv:2603.22799 [cs].
- Lester James Validad Miranda, Elyanah Aco, Conner G. Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. [FilBench: Can LLMs Understand and Generate Filipino?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2496–2529, Suzhou, China. Association for Computational Linguistics.
- Grace Muzny and Luke Zettlemoyer. 2013. [Automatic idiom identification in wiktionary](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421. Association for Computational Linguistics.
- Hiroki Nakayama. 2018. [sequeval: A python framework for sequence label evaluation](#). GitHub repository. Accessed: 2026-05-18.
- OpenAI. 2022. [ChatGPT](#). Accessed: 2026-05-24.
- OpenAI. 2024. [GPT-4o mini: Advancing cost-efficient intelligence](#). Accessed: 2026-05-24.
- Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. [Verbal multiword expression identification: Do we need a sledgehammer to crack a nut?](#) In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Juan Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. [Mind the \(Language\) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts](#).
- Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. [Chinese idiom paraphrasing](#). *Transactions of the Association for Computational Linguistics*, 11:740–754.
- Giancarlo Salton, Robert Ross, and John Kelleher. 2014. [An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese](#). In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden. Association for Computational Linguistics.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. [FaceNet: A unified embedding for face recognition and clustering](#). pages 815–823. IEEE Computer Society.
- Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2023. [RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting](#). *arXiv preprint*. ArXiv:2305.15685 [cs].
- Atakan Site, Oğuz Ali Arslan, and Gülşen Eryiğit. 2026. [ITUNLP at MWE-2026 AdMIRE 2: A Zero-Shot LLM Pipeline for Multimodal Idiom Understanding and Ranking](#). In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 226–236, Rabat, Morocco. Association for Computational Linguistics.
- Gian Carlos Tan, Jhan Kyle Canlas, Ren Joseph Ayangco, Daeschlan Blane Gador, Mico Magtira, Jean Malolos, Ramon Rodriguez, Joseph Marvin Imperial, and Mideth Abisado. 2024. [CebBERT: A Lightweight Data-Transparent DistilBERT Model for Cebuano Language Processing](#). In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 904–913, Tokyo, Japan. Tokyo University of Foreign Studies.
- Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. [ID10M: Idiom Identification in 10 Languages](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. [Unsupervised paraphrasing of multiword expressions](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4732–4746, Toronto, Canada. Association for Computational Linguistics.
- Wiktionary contributors. 2016. [Category:Cebuano idioms - Wiktionary, the free dictionary](#).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengqiang Xu, Chuanqi Chen, Chao Jiang, Dajie Dong, and 1 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.

Ziheng Zeng and Suma Bhat. 2021. [Idiomatic Expression Identification using Semantic Compatibility](#). *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2022. [Idiomatic expression paraphrasing without strong supervision](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11774–11782.

A Reproducibility and Code Availability

The CebIDN code, dataset, model weights, evaluation scripts, prompts, and human annotation results are publicly available at <https://github.com/DaRainFlavor/CebIDN>. The code is released under MIT, and the dataset under CC BY-SA 4.0.

B Human Evaluation and Rating Distributions

Evaluators in all three human studies were native-level Cebuano speakers recruited voluntarily through local networks in Cebu. The studies comprised: (1) a 100-sentence dataset quality evaluation of figurative-paraphrase adequacy and fluency; (2) a 136-sentence category classification task; and (3) a 100-sentence candidate normalization evaluation for validating the LLM-as-a-Judge framework. Participants included students, public school teachers, and other native speakers. No financial compensation was provided, and participation was consensual.

For the 136-sentence classification task, three reviewers classified each sentence as *Figurative (Metaphorical)*, *Literal*, or *Hard Negative*, given the target idiom and its literal and figurative definitions. The results are reported in Section 4.2.

For the dataset quality evaluation, two figurative-paraphrase pairs per idiom (100 pairs total) were independently rated on 5-point Likert scales. Mean ratings were 4.96 ± 0.19 for Adequacy (Semantic Fidelity) and 4.93 ± 0.29 for Fluency, with pairwise agreement of 92.67

Fleiss’ κ was -0.038 for Adequacy and 0.104 for Fluency. The low κ values despite high percentage agreement reflect the strong ceiling effect in the ratings, with most scores concentrated at 5. Figure B.1 shows the rating distributions.

LLM-as-a-Judge validation, reported in the main text, yielded Spearman correlations of $\rho = 0.791$, 0.690 , and 0.498 with human ratings for Semantic Fidelity, Span Accuracy, and Fluency, respectively.

B.1 Annotation Guidelines and Rubrics

The native Cebuano evaluators used the following rubrics through the custom web interface.

Adequacy (Semantic Fidelity):

- **5:** Perfect preservation of meaning.
- **4:** Most meaning preserved; minor nuances lost.
- **3:** Core meaning remains, but noticeably altered.
- **2:** Only fragments of meaning remain.
- **1:** Completely different or nonsensical.

Fluency:

- **5:** Perfectly natural, grammatically correct, and free of spelling/typographical errors.
- **4:** Good Cebuano; slight awkwardness or minor spelling/punctuation errors, but acceptable.
- **3:** Understandable, but contains noticeable grammatical errors or frequent spelling mistakes that distract the reader.
- **2:** Very unnatural; machine-generated feel.
- **1:** Incomprehensible or completely broken.

Additional Guidelines:

- Evaluators assessed each original-paraphrase pair together for semantic preservation.
- Idiom definitions were specific to the contexts used in the dataset and might not cover all lexical meanings.
- Evaluators were instructed not to penalize conversational or informal Cebuano solely for stylistic shortcuts.

C Hyperparameter and Sweep Details

This section outlines the hyperparameter settings for Stage 1 (CebBERT fine-tuning) and Stage 2 (Qwen-2.5-7B fine-tuning) along with the validation sweep for the contrastive loss weight. All experiments were conducted using Kaggle’s academic compute infrastructure equipped with dual NVIDIA T4 GPUs. The total computational budget across all training, hyperparameter sweeps, and evaluations was approximately 43 GPU hours. Specifically, Stage 1 sequence labeling (including the parameter sweep and three random seeds) required approximately 10 GPU hours. Stage 2 gen-

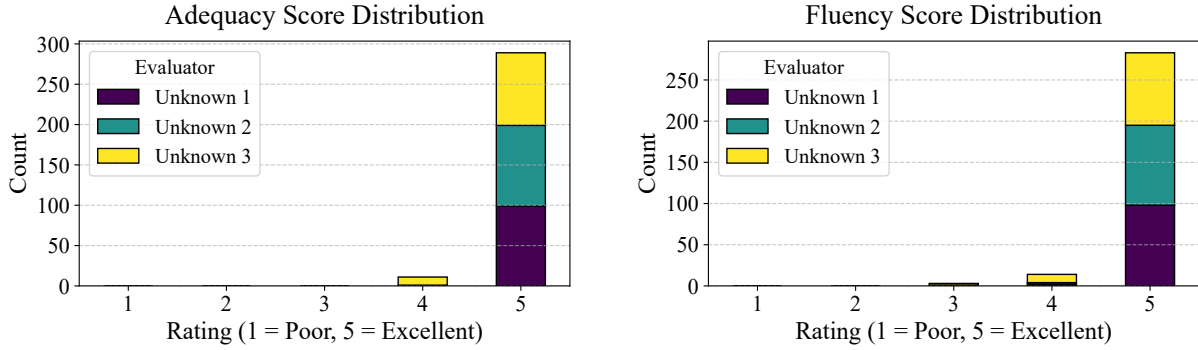


Figure B.1: Rating score distributions from the three native Cebuano evaluators for adequacy (left) and fluency (right), highlighting the strong ceiling effect in the ratings.

erative fine-tuning, retrieval-augmented baselines, and evaluations required approximately 33 GPU hours (roughly 11 hours per seed, conducted in staged training intervals with intermediate validation and postprocessing checkpoints).

The models used in this work comprise CebBERT ($\sim 66\text{M}$ parameters) and Qwen-2.5-7B ($\sim 7.6\text{B}$ parameters). Table C.1 and Table C.2 present the parameter details for the first and second stages, respectively.

C.1 Stage 1: CebBERT Fine-Tuning

The Stage 1 model used CebBERT as the encoder backbone. To preserve label consistency for CRF decoding, subword tokens inherited the BIO label of the originating word token. Hyperparameters used for fine-tuning are listed in Table C.1.

Hyperparameter	Value
Encoder learning rate	2×10^{-5}
Classifier learning rate	1×10^{-3}
Batch size	16
Max epochs	15
Max sequence length	128 tokens
Dropout rate	0.1
Triplet margin (m)	1.0
Early stopping patience	3 epochs

Table C.1: Stage 1 hyperparameters (CebBERT fine-tuning).

C.2 Stage 2: Qwen-2.5-7B Fine-Tuning

Stage 2 fine-tuned Qwen-2.5-7B-Instruct with 4-bit NF4 quantization and LoRA. Joint training was conducted on knowledge-augmented prompts (hints + tags) and tag-only prompts to improve robustness when retrieved metadata was unavailable. For the end-to-end baseline, figurative samples were upsampled to reduce class imbalance. Hyperparameters are listed in Table C.2.

Hyperparameter	Value
Learning rate	2×10^{-4}
Effective batch size	16
Max epochs	5
Max sequence length	512 tokens
Quantization	4-bit (NF4)
LoRA rank (r)	16
LoRA alpha	16
Warmup steps	10
Early stopping patience	3 epochs

Table C.2: Stage 2 hyperparameters (Qwen-2.5-7B fine-tuning).

C.3 Contrastive Weight Alpha Sweep

To determine the optimal contrastive weight α in Equation 4, a hyperparameter sweep was performed over $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$ using the proposed BIO+CRF+TL configuration on Seed 42. Table C.3 shows the average loss and validation F1 score (for idiom spans) across the different values. The sweep showed that $\alpha = 0.05$ yielded the lowest average loss (0.1642) and the highest validation F1 score (0.8998).

Alpha (α)	Avg Loss	Val F1 (Idiom)
0.01	0.2083	0.8832
0.05	0.1642	0.8998
0.10	0.2256	0.8915
0.20	0.5353	0.8789
0.50	0.5353	0.8605

Table C.3: Hyperparameter sweep results for contrastive loss weight α evaluated on the validation set using the BIO+CRF+TL configuration (Seed 42). The selected optimal value is highlighted in bold.

D Stage 1 Expanded Identification Breakdown

Table D.1 provides the full Precision, Recall, F1, and Partial breakdown for all Stage 1 configurations evaluating idiom-only triggers. In addi-

tion, Table D.2 shows the strict model sequence-level match performance evaluating the complete tagging schemes (both figurative and literal tags). Lastly, Table D.3 presents the detailed Precision, Recall, F1, and Accuracy breakdown for the binary decision task (idiom vs. literal disambiguation). Figure D.1 plots the inverse relationship between boundary precision and Trigger F1 across configurations.

E Expanded Normalization Quality Breakdowns

This section provides the complete multidimensional breakdown of our normalization evaluation. Table E.1 compares performance across the few-shot prompting and fine-tuned settings on all Stage 2 configurations. Table E.2 details the end-to-end normalization quality when conditional on successful Stage 1 trigger identification. Finally, Table E.3 evaluates normalization resilience on figurative sentences with imperfectly predicted spans, comparing results with and without canonical recovery from the Idiom Knowledge Base (IKB). In addition to the metrics described in the main text, we report ROUGE-L in these tables as an additional measure of longest common subsequence overlap.

F End-to-End Comparison with Tagged Normalization

This section presents the end-to-end system comparison using the FT+hints+tags for with canonical match and FT+tags as fallback. (Table F.1). In this configuration, the Stage 2 normalizer is prompted with explicit boundary tags (`<idiom> . . . </idiom>`) to mark the detected idiom span. As discussed in Section 3.3, while this configuration enforces strict span-level editing constraints and achieves lower latency, it is sensitive to Stage 1 boundary errors, which propagate downstream and degrade translation quality compared to the tag-free normalizer configuration used in Table 6.

G Calculation of Overall Pipeline Latency

This section details the overall end-to-end latency calculation for the modular CebIDN configurations reported in Tables 6 and F.1.

For a dataset of N total sentences, the total processing time T_{total} is computed as:

$$T_{\text{total}} = N \cdot L_{\text{tag}} + N_D \cdot L_{\text{norm,hint}} + N_C \cdot L_{\text{norm,nohint}} \quad (5)$$

where:

- L_{tag} is the average inference latency of the Stage 1 sequence labeler (CebBERT) per sentence.
- N_D is the number of sentences where Stage 1 predicted an idiom span and a matching canonical entry was successfully retrieved from the IKB, triggering a Stage 2 prompt with hints.
- $L_{\text{norm,hint}}$ is the average inference latency of the Stage 2 generative normalizer (Qwen-2.5-7B) when provided with retrieved IKB hints.
- N_C is the number of sentences where Stage 1 predicted an idiom span but no matching canonical entry was found in the IKB, triggering a Stage 2 prompt without hints.
- $L_{\text{norm,nohint}}$ is the average inference latency of the Stage 2 generative normalizer without retrieved hints.

Liter sentences that are correctly classified by Stage 1 as containing no idiomatic spans bypass Stage 2 entirely, incurring zero generative latency.

The overall average latency per sentence L_{E2E} is then calculated as the total time divided by the total number of sentences:

$$\begin{aligned} L_{\text{E2E}} &= \frac{T_{\text{total}}}{N} \\ &= L_{\text{tag}} + \frac{1}{N} (N_D \cdot L_{\text{norm,hint}} \\ &\quad + N_C \cdot L_{\text{norm,nohint}}) \end{aligned} \quad (6)$$

H Prompt Templates

This section presents the prompt templates used in our experiments. The five core template structures are shown in Figures H.1–H.5; the remaining configurations are derived from these templates using the rules below. Table H.1 provides English translations of the Cebuano in-context exemplars for reader convenience; these idioms are not among the 50 dataset idioms, and the translations were not included in the prompts.

Zero-Shot (ZS) and End-to-End Zero-Shot (E2E ZS) configurations are obtained from their Few-Shot counterparts (Figures H.2 and H.1) by removing all in-context examples, leaving only the system prompt and final input.

Tagged configurations are derived from their tag-free counterparts by wrapping the predicted idiom span with `<idiom></idiom>` and instructing the model to replace the tagged phrase with its literal equivalent. Few-shot exemplars and training targets reinforce that the final output contains no tags.

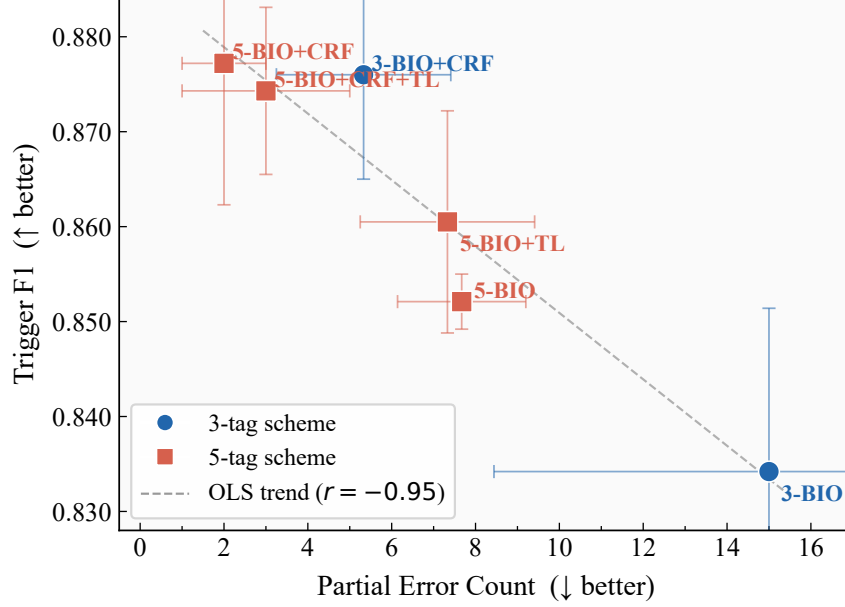


Figure D.1: Partial Match count vs. Trigger F1 across Stage 1 configurations, grouped by tag scheme. Error bars denote standard deviation over three seeds. Dashed line: OLS regression trend ($r = -0.95$).

Config	Metrics			
	Trigger F1	Precision	Recall	Partial
3-tag				
BIO	0.8342 $\pm$ 0.0170	0.7988 $\pm$ 0.0330	0.8733 $\pm$ 0.0020	15.00 $\pm$ 6.60
BIO+CRF	0.8760 $\pm$ 0.0110	0.8582 $\pm$ 0.0150	0.8947 $\pm$ 0.0120	5.33 $\pm$ 2.10
5-tag				
BIO	0.8521 $\pm$ 0.0030	0.8377 $\pm$ 0.0260	0.8680 $\pm$ 0.0220	7.67 $\pm$ 1.50
BIO+CRF	0.8772 $\pm$ 0.0150	0.8589 $\pm$ 0.0390	0.8973 $\pm$ 0.0140	2.00 $\pm$ 1.00
BIO+TL	0.8605 $\pm$ 0.0120	0.8168 $\pm$ 0.0380	0.9107 $\pm$ 0.0210	7.33 $\pm$ 2.10
BIO+CRF+TL	0.8743 $\pm$ 0.0090	0.8702 $\pm$ 0.0370	0.8800 $\pm$ 0.0310	3.00 $\pm$ 2.00

Table D.1: Expanded Stage 1 (CebBERT, Table 3) idiom identification results. TL denotes Triplet Loss. Best results within each metric in every tagging scheme are highlighted in bold.

Conversely, FS + hints is obtained from FS + hints + tags (Figure H.3) by removing the boundary tags.

E2E RAG configurations are derived from the E2E Few-Shot template (Figure H.1) by inserting the dictionary metadata block from Figure H.5 above the input sentence; the fine-tuned RAG variant omits the few-shot exemplars.

I LLM-as-Judge Evaluation Prompt and Rubric

This section presents the prompt template and grading rubric used for our LLM-as-a-judge evaluation. To compute the automatic quality scores shown in Table J.1 and Table K.1, we use GPT-4o-mini to grade normalization candidates across three dimensions including Semantic Fidelity, Span Accuracy, and Fluency. The exact prompt template, including the detailed grading rubric and JSON response constraints, is shown in Figure I.1.

J Baseline Normalization Failures

This appendix presents qualitative examples of normalization errors generated by untuned baselines when dictionary hints and tags are absent. Table J.1 categorizes these failures into distinct error types, showing how the lack of external resources degrades performance.

While showing the results for all other configurations would make the table too long, this also applies to end-to-end Few Shot and Zero Shot these failures are resolved once hints or tuning are applied. Augmenting the normalizer with retrieved metadata or model fine-tuning successfully maps these inputs to the expected translations.

Config	Metrics				
	F1	Precision	Recall	Accuracy	Inv BIO
3-tag					
BIO	0.8342 $\pm$ 0.0172	0.7988 $\pm$ 0.0332	0.8733 $\pm$ 0.0023	0.8976 $\pm$ 0.0112	16.00 $\pm$ 7.81
BIO+CRF	0.8760 $\pm$ 0.0110	0.8582 $\pm$ 0.0154	0.8947 $\pm$ 0.0115	0.9162 $\pm$ 0.0074	5.67 $\pm$ 2.89
5-tag					
BIO	0.8042 $\pm$ 0.0107	0.7628 $\pm$ 0.0095	0.8503 $\pm$ 0.0122	0.8593 $\pm$ 0.0074	27.67 $\pm$ 5.51
BIO+CRF	0.8347 $\pm$ 0.0257	0.8244 $\pm$ 0.0237	0.8452 $\pm$ 0.0278	0.8750 $\pm$ 0.0195	9.67 $\pm$ 6.43
BIO+TL	0.8069 $\pm$ 0.0097	0.7660 $\pm$ 0.0148	0.8526 $\pm$ 0.0064	0.8583 $\pm$ 0.0052	32.33 $\pm$ 2.52
BIO+CRF+TL	0.8342 $\pm$ 0.0068	0.8082 $\pm$ 0.0187	0.8622 $\pm$ 0.0080	0.8686 $\pm$ 0.0112	10.33 $\pm$ 2.89

Table D.2: Stage 1 strict model performance (exact sequence match) comparison across different CebBERT tagging configurations. TL denotes Triplet Loss. F1, Precision, and Recall are entity-level metrics calculated using sequeval on the complete tag schema (evaluating both Figurative and Literal spans in the 5-tag version, and Figurative only in the 3-tag version). Accuracy measures the percentage of sentences with a perfect sequence-level match. Inv BIO counts predicted sequences with illegal BIO transitions. Best results within each metric in every tagging scheme are highlighted in bold.

Config	Metrics			
	F1	Precision	Recall	Accuracy
3-tag				
BIO	0.8941 $\pm$ 0.0008	0.8676 $\pm$ 0.0172	0.9227 $\pm$ 0.0201	0.9196 $\pm$ 0.0017
BIO+CRF	0.8986 $\pm$ 0.0085	0.8832 $\pm$ 0.0188	0.9147 $\pm$ 0.0023	0.9240 $\pm$ 0.0073
5-tag				
BIO	0.8860 $\pm$ 0.0046	0.8775 $\pm$ 0.0321	0.8960 $\pm$ 0.0250	0.9152 $\pm$ 0.0059
BIO+CRF	0.8859 $\pm$ 0.0128	0.8696 $\pm$ 0.0366	0.9040 $\pm$ 0.0144	0.9142 $\pm$ 0.0123
BIO+TL	0.8933 $\pm$ 0.0160	0.8566 $\pm$ 0.0409	0.9347 $\pm$ 0.0167	0.9176 $\pm$ 0.0147
BIO+CRF+TL	0.8903 $\pm$ 0.0132	0.8912 $\pm$ 0.0406	0.8907 $\pm$ 0.0244	0.9191 $\pm$ 0.0117
E2E LLM				
Zero-Shot	0.4593 $\pm$ 0.0000	0.3570 $\pm$ 0.0000	0.6440 $\pm$ 0.0000	0.4426 $\pm$ 0.0000
Few-Shot (7-shot)	0.4276 $\pm$ 0.0223	0.4507 $\pm$ 0.0265	0.4067 $\pm$ 0.0189	0.5995 $\pm$ 0.0180
Fine-Tuned	0.8470 $\pm$ 0.0155	0.8542 $\pm$ 0.0671	0.8480 $\pm$ 0.0779	0.8877 $\pm$ 0.0110
E2E RAG				
Few-Shot RAG (7-shot)	0.5438 $\pm$ 0.0031	0.3832 $\pm$ 0.0031	0.9360 $\pm$ 0.0000	0.4225 $\pm$ 0.0073
Fine-Tuned RAG	0.8685 $\pm$ 0.0223	0.8591 $\pm$ 0.0073	0.8787 $\pm$ 0.0411	0.9025 $\pm$ 0.0147

Table D.3: Expanded Binary Decision Performance (Idiom vs. Literal Classification, from Table 6) comparing sequence labeling (CebBERT) and generative (Qwen-2.5-7B) configurations. TL denotes Triplet Loss. RAG Baseline is oracle given hints of the idiom used and it is up to the model to normalize or not. Best results per metric within each group are highlighted in bold.

K XML Tag Leakage and Fragility of Tagged Normalization

This section analyzes the qualitative formatting and translation errors observed under tagged normalization configurations when processing imperfect boundary inputs from Stage 1. Table K.1 illustrates three distinct failure modes: tag leakage, tag shifting, and boundary truncation leading to crossover hallucinations.

L Qualitative Error Analysis: Idiom Types, Morphology, and Error Propagation

This section details the qualitative and quantitative analysis of system errors across idiom types, mor-

phological variations, and downstream propagation modes.

Literal-Figurative Ambiguity by Idiom Type

Table L.1 reports the performance of the optimal 3-tag BIO+CRF configuration across the three idiom classes, averaged over three evaluation seeds. Complete idioms are the most challenging, with the lowest detection F1 (87.10%) and Composite score (4.61), whereas No-Literal idioms achieve 98.71% detection F1 and 4.98 Composite.

Agglutinative Morphological Variations Cebuano’s rich agglutinative morphology causes boundary detection failures, particularly when inflected verb forms obscure canonical idiom spans. Under the optimal 3-tag BIO+CRF Stage 1 con-

Config	Automatic Metrics			LLM-as-Judge			
	BLEU	ROUGE-L	BERTScore	Semantic	Span	Fluency	Composite
ZS	48.62 $\pm$ 0.00	69.90 $\pm$ 0.00	76.94 $\pm$ 0.00	1.04 $\pm$ 0.00	1.02 $\pm$ 0.00	3.86 $\pm$ 0.00	1.98 $\pm$ 0.00
FS	53.83 $\pm$ 0.22	73.45 $\pm$ 0.20	79.26 $\pm$ 0.06	1.09 $\pm$ 0.01	1.05 $\pm$ 0.01	3.87 $\pm$ 0.02	2.00 $\pm$ 0.01
FS + hints	73.16 $\pm$ 0.63	86.91 $\pm$ 0.39	90.38 $\pm$ 0.15	3.39 $\pm$ 0.07	3.27 $\pm$ 0.07	4.32 $\pm$ 0.03	3.66 $\pm$ 0.06
FS + hints + tags	76.53 $\pm$ 0.64	88.75 $\pm$ 0.29	91.88 $\pm$ 0.20	3.88 $\pm$ 0.03	3.74 $\pm$ 0.04	4.42 $\pm$ 0.01	4.01 $\pm$ 0.03
FT	86.10 $\pm$ 0.63	92.56 $\pm$ 0.61	94.57 $\pm$ 0.49	4.30 $\pm$ 0.08	4.25 $\pm$ 0.08	4.60 $\pm$ 0.06	4.39 $\pm$ 0.07
FT + hints	88.51 $\pm$ 0.25	94.89 $\pm$ 0.16	96.37 $\pm$ 0.03	4.71 $\pm$ 0.02	4.65 $\pm$ 0.02	4.85 $\pm$ 0.01	4.74 $\pm$ 0.01
FT + tags	86.00 $\pm$ 0.84	92.80 $\pm$ 0.51	94.81 $\pm$ 0.42	4.34 $\pm$ 0.02	4.29 $\pm$ 0.01	4.63 $\pm$ 0.03	4.42 $\pm$ 0.01
FT + hints + tags	88.51 $\pm$ 0.59	94.58 $\pm$ 0.32	96.18 $\pm$ 0.06	4.67 $\pm$ 0.01	4.60 $\pm$ 0.02	4.80 $\pm$ 0.02	4.69 $\pm$ 0.02

Table E.1: Expanded performance comparison across Few-Shot and Fine-Tuned Normalization settings using Qwen-2.5-7B. Best results within each metric per group are highlighted in bold.

Config	Automatic Metrics			LLM-as-Judge			
	BLEU	ROUGE-L	BERTScore	Semantic	Span	Fluency	Composite
3-tag + FT+Hints and FT Fallback Normalizer							
BIO	88.85 $\pm$ 0.57	94.64 $\pm$ 0.35	96.33 $\pm$ 0.20	4.67 $\pm$ 0.04	4.60 $\pm$ 0.04	4.83 $\pm$ 0.01	4.70 $\pm$ 0.03
BIO+CRF	88.87 $\pm$ 0.51	94.74 $\pm$ 0.39	96.40 $\pm$ 0.12	4.69 $\pm$ 0.03	4.62 $\pm$ 0.04	4.84 $\pm$ 0.01	4.71 $\pm$ 0.03
5-tag + FT+Hints and FT Fallback Normalizer							
BIO	89.12 $\pm$ 0.68	94.87 $\pm$ 0.37	96.47 $\pm$ 0.17	4.70 $\pm$ 0.03	4.63 $\pm$ 0.04	4.85 $\pm$ 0.01	4.73 $\pm$ 0.03
BIO+CRF	89.13 $\pm$ 0.25	94.86 $\pm$ 0.16	96.53 $\pm$ 0.06	4.71 $\pm$ 0.02	4.64 $\pm$ 0.02	4.85 $\pm$ 0.01	4.74 $\pm$ 0.02
BIO+TL	88.87 $\pm$ 0.53	94.76 $\pm$ 0.25	96.39 $\pm$ 0.08	4.69 $\pm$ 0.02	4.62 $\pm$ 0.03	4.84 $\pm$ 0.00	4.72 $\pm$ 0.02
BIO+CRF+TL	89.23 $\pm$ 0.25	94.88 $\pm$ 0.18	96.49 $\pm$ 0.02	4.70 $\pm$ 0.02	4.63 $\pm$ 0.01	4.84 $\pm$ 0.01	4.73 $\pm$ 0.01
E2E LLM							
ZS	53.02 $\pm$ 0.00	73.16 $\pm$ 0.00	79.44 $\pm$ 0.00	1.09 $\pm$ 0.00	1.06 $\pm$ 0.00	4.25 $\pm$ 0.00	2.13 $\pm$ 0.00
FS	54.35 $\pm$ 0.68	74.07 $\pm$ 0.45	79.51 $\pm$ 0.52	1.03 $\pm$ 0.03	1.02 $\pm$ 0.03	4.20 $\pm$ 0.06	2.08 $\pm$ 0.04
FT	86.89 $\pm$ 0.89	93.22 $\pm$ 0.53	95.04 $\pm$ 0.40	4.39 $\pm$ 0.06	4.34 $\pm$ 0.05	4.65 $\pm$ 0.02	4.46 $\pm$ 0.04
RAG (Oracle)							
FS	74.18 $\pm$ 0.29	87.43 $\pm$ 0.09	90.50 $\pm$ 0.16	3.62 $\pm$ 0.05	3.50 $\pm$ 0.06	4.43 $\pm$ 0.03	3.85 $\pm$ 0.05
FT	89.21 $\pm$ 0.66	94.93 $\pm$ 0.35	96.39 $\pm$ 0.21	4.69 $\pm$ 0.00	4.62 $\pm$ 0.02	4.84 $\pm$ 0.02	4.72 $\pm$ 0.02

Table E.2: Expanded Stage 2 Idiom Normalization Quality Evaluation. Few-shot (FS) configurations use 7 in-context examples. All configurations utilize Qwen-2.5-7B. Comparison of normalization performance for sentences where an idiom trigger was successfully identified by the Stage 1 sequence labeler (CebBERT). Pipeline configurations (3-tag and 5-tag) utilize adaptive routing (FT+ hints when a canonical match is retrieved from the dictionary, falling back to FT otherwise). Quality is measured through semantic fidelity, span accuracy, and fluency, evaluated via both automatic metrics (BLEU, ROUGE-L, BERTScore) and a multi-dimensional LLM-as-a-Judge framework. Best results within each metric per group are highlighted in bold.

figuration, morphological inflection (where base words are modified by affixation) led to an average strict sequence matching failure rate of 20.63% $\pm$ 9.91% (representing 13 failures out of 63 total cases across three evaluation seeds). For example, inflected verb forms like *maghingaon* (from the base idiom *hingaon og nangka*) or *niagi* (from *agi sa bangag sa dagom*) shift expected boundaries, leading the sequence labeler to either truncate the span or omit the trigger entirely.

Failure Propagation and Hybrid Rewrites

Most partial-span detection errors cannot be matched to the IKB morphology-aware substring matching rules, preventing the retrieval of dictionary definitions and hints. This produces hybrid literal-figurative Cebuano rewrites or syntax hallu-

inations (as documented in Table J.1).

M Pipeline Normalization Success Case Studies

This section presents detailed qualitative case studies demonstrating how the CebIDN optimal pipeline configurations (3-tag BIO+CRF with tagged normalizer and 3-tag BIO with tag-free normalizer) successfully process different idiom detection scenarios. We outline cases representing perfect span detection, imperfect span detection with canonical recovery, and imperfect span detection with fallback routing across both tagged and tag-free normalizers.

Config	Automatic Metrics			LLM-as-Judge			
	BLEU	ROUGE-L	BERTScore	Semantic	Span	Fluency	Composite
With Canonical							
ZS	58.96 $\pm$ 0.00	78.43 $\pm$ 0.00	81.41 $\pm$ 0.00	1.21 $\pm$ 0.00	1.14 $\pm$ 0.00	4.00 $\pm$ 0.00	2.12 $\pm$ 0.00
FS	56.42 $\pm$ 0.40	78.26 $\pm$ 0.15	80.54 $\pm$ 0.18	1.47 $\pm$ 0.00	1.35 $\pm$ 0.00	3.37 $\pm$ 0.07	2.07 $\pm$ 0.02
FS + hints	76.86 $\pm$ 0.41	90.41 $\pm$ 0.38	92.12 $\pm$ 0.20	3.71 $\pm$ 0.21	3.50 $\pm$ 0.26	4.45 $\pm$ 0.08	3.89 $\pm$ 0.18
FS + hints + tags	77.88 $\pm$ 0.81	89.53 $\pm$ 0.34	91.88 $\pm$ 0.75	3.69 $\pm$ 0.17	3.59 $\pm$ 0.20	4.20 $\pm$ 0.09	3.82 $\pm$ 0.15
FT	85.98 $\pm$ 2.05	93.66 $\pm$ 1.57	94.10 $\pm$ 1.04	3.93 $\pm$ 0.40	3.86 $\pm$ 0.40	4.33 $\pm$ 0.25	4.04 $\pm$ 0.35
FT + hints	90.03 $\pm$ 1.50	95.12 $\pm$ 1.07	96.58 $\pm$ 0.33	4.74 $\pm$ 0.04	4.60 $\pm$ 0.11	4.86 $\pm$ 0.07	4.73 $\pm$ 0.07
FT + tags	70.29 $\pm$ 7.57	87.09 $\pm$ 3.66	86.01 $\pm$ 5.22	2.75 $\pm$ 0.32	2.61 $\pm$ 0.34	3.47 $\pm$ 0.36	2.94 $\pm$ 0.34
FT + hints + tags	77.08 $\pm$ 2.77	91.63 $\pm$ 1.27	90.51 $\pm$ 2.44	3.57 $\pm$ 0.15	3.45 $\pm$ 0.24	4.04 $\pm$ 0.15	3.69 $\pm$ 0.18
Without Canonical							
ZS	45.03 $\pm$ 0.00	66.07 $\pm$ 0.00	75.29 $\pm$ 0.00	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	3.76 $\pm$ 0.00	1.92 $\pm$ 0.00
FS	50.94 $\pm$ 1.07	71.05 $\pm$ 0.62	78.11 $\pm$ 0.63	1.03 $\pm$ 0.03	1.03 $\pm$ 0.03	3.84 $\pm$ 0.20	1.97 $\pm$ 0.07
FT	83.50 $\pm$ 2.28	90.33 $\pm$ 2.37	92.54 $\pm$ 1.31	3.82 $\pm$ 0.19	3.78 $\pm$ 0.17	4.32 $\pm$ 0.13	3.98 $\pm$ 0.16
FT + tags	67.17 $\pm$ 2.53	80.98 $\pm$ 0.93	86.11 $\pm$ 0.74	2.57 $\pm$ 0.05	2.51 $\pm$ 0.06	3.78 $\pm$ 0.06	2.95 $\pm$ 0.02

Table E.3: Expanded performance comparison on predicted imperfect idiom spans from Stage 1 (CebBERT). Pipeline configurations utilize adaptive routing (FT+hints when a canonical match is recovered, falling back to FT otherwise). Few-shot (FS) configurations use 3 in-context examples. All configurations utilize Qwen-2.5-7B. Subsets are partitioned based on whether the canonical idiom was recoverable from the predicted span. Best results within each metric per group are highlighted in bold.

Config	Bin. F1	BLEU	BERTScore	Composite	Avg. Quality	Speed (ms/sent)
3-tag BIO	0.894 $\pm$ 0.001	87.89 $\pm$ 0.31	95.75 $\pm$ 0.26	4.61 $\pm$ 0.04	89.08 $\pm$ 0.37	475 $\pm$ 91
3-tag BIO+CRF	0.899 $\pm$ 0.009	88.27 $\pm$ 0.30	96.05 $\pm$ 0.16	4.66 $\pm$ 0.01	89.74 $\pm$ 0.74	468 $\pm$ 109
5-tag BIO	0.886 $\pm$ 0.005	87.50 $\pm$ 0.53	95.80 $\pm$ 0.03	4.63 $\pm$ 0.02	88.60 $\pm$ 0.64	464 $\pm$ 127
5-tag BIO+CRF	0.886 $\pm$ 0.013	88.41 $\pm$ 0.31	96.11 $\pm$ 0.19	4.67 $\pm$ 0.01	88.83 $\pm$ 1.26	469 $\pm$ 97
5-tag BIO+TL	0.893 $\pm$ 0.016	87.76 $\pm$ 0.43	95.81 $\pm$ 0.19	4.63 $\pm$ 0.03	88.67 $\pm$ 1.33	487 $\pm$ 79
5-tag BIO+CRF+TL	0.890 $\pm$ 0.013	88.40 $\pm$ 0.47	96.00 $\pm$ 0.14	4.65 $\pm$ 0.04	89.18 $\pm$ 1.13	449 $\pm$ 70
E2E ZS	0.459 $\pm$ 0.000	53.02 $\pm$ 0.00	79.44 $\pm$ 0.00	2.13 $\pm$ 0.00	33.28 $\pm$ 0.00	551 $\pm$ 1
E2E FS	0.428 $\pm$ 0.022	54.35 $\pm$ 0.68	79.51 $\pm$ 0.52	2.08 $\pm$ 0.04	53.02 $\pm$ 1.46	1166 $\pm$ 206
E2E FT	0.847 $\pm$ 0.016	86.89 $\pm$ 0.89	95.04 $\pm$ 0.40	4.46 $\pm$ 0.04	85.48 $\pm$ 1.21	924 $\pm$ 214
RAG (Oracle) FS	0.544 $\pm$ 0.003	74.18 $\pm$ 0.29	90.50 $\pm$ 0.16	3.85 $\pm$ 0.05	34.89 $\pm$ 0.83	1666 $\pm$ 197
RAG (Oracle) FT	0.869 $\pm$ 0.022	89.21 $\pm$ 0.66	96.39 $\pm$ 0.21	4.72 $\pm$ 0.02	87.93 $\pm$ 1.35	929 $\pm$ 118

Table F.1: End-to-end system comparison using standard tagged routing, comparing sequence labeling (CebBERT) and generative (Qwen-2.5-7B) configurations. CebIDN configurations utilize FT+hints+tags when a canonical match is retrieved from the Idiomatic Knowledge Base (IKB), and fall back to FT+tags when no canonical match is found. Best results per metric are highlighted in bold.

Idiom	Cebuano Exemplar Sentence	English Translation
<i>katagakon ang mata</i>	Katagakon na akong mata, ugma nalang ni nato tiwason. (Figurative Input Exemplar)	My eyes are about to fall out, let us finish this tomorrow. [Literal Translation]
	Katugon na ko, ugma nalang ni nato tiwason. (Output Exemplar / Paraphrase)	I am sleepy, let us finish this tomorrow.
	Dako kaayo og mata si Elsa, mura og katagakon ang iyang mata. (Literal Use Exemplar)	Elsa's eyes are very big, they look like they are about to fall out.
	Bantayi ang bata kay matagak unya sa hagdan. (No Idiom Phrase Negative Exemplar)	Watch the child because he/she might fall from the stairs.
<i>lapas sa kalendaryo</i>	Kadugay jud ninyo nagpakasal? Lapas na man mo sa kalendaryo! (Figurative Input Exemplar)	Why did it take you so long to get married? You are past the calendar! [Literal Translation]
	Kadugay jud ninyo nagpakasal? Lapas na man mo sa trayntay uno nga pangedaron! (Output Exemplar / Paraphrase)	Why did it take you so long to get married? You are past thirty-one years of age!
	Tan-awa daw sa kalendaryo kung naa bay holiday karong bulana. (No Idiom Phrase Negative Exemplar)	Please look at the calendar if there is any holiday this month.
<i>nalumos sa sabaw</i>	Nihilak ang transferee kay gisungog nga nalumos daw siya sa sabaw. (Figurative Input Exemplar)	The transferee cried because he/she was teased for drowning in soup. [Literal Translation]
	Nihilak ang transferee kay gisungog nga wala daw siyay amahan. (Output Exemplar / Paraphrase)	The transferee cried because he/she was teased for not having a father.
	Tuga-tuga og langoy, hapit nuon nalumos! (No Idiom Phrase Negative Exemplar)	Daring to swim, you almost drowned!

Table H.1: English translations of Cebuano in-context exemplar sentences used in the few-shot prompt templates.

<pre> < im_start >system You are a Cebuano language expert. Analyze the sentence provided. 1. Determine if it contains an idiom used figuratively (metaphorically). 2. If YES: Output ONLY the entire sentence with the idiomatic part replaced by its literal Cebuano equivalent. 3. If NO (no idiom or idiom used literally): Output ONLY the word "None". CRITICAL: Provide ONLY the direct answer. Do NOT include explanations, "Explanation:", or any other text.< im_end > < im_start >user Katagakon na akong mata, ugma nalang ni nato tiwason.< im_end > < im_start >assistant Katugon na ko, ugma nalang ni nato tiwason.< im_end > < im_start >user Dako kaayo og mata si Elsa, mura og katagakon ang iyang mata.< im_end > < im_start >assistant None< im_end > < im_start >user Bantayi ang bata kay matagak unya sa hagdan.< im_end > < im_start >assistant None< im_end > < im_start >user Kadugay jud ninyo nagpakasal? Lapas na man mo sa kalendaryo!< im_end > < im_start >assistant Kadugay jud ninyo nagpakasal? Lapas na man mo sa trayntay uno nga pangedaron!< im_end > < im_start >user Tan-awa daw sa kalendaryo kung naa bay holiday karong bulana.< im_end > < im_start >assistant None< im_end > < im_start >user Nihilak ang transferee kay gisungog nga nalumos daw siya sa sabaw.< im_end > < im_start >assistant Nihilak ang transferee kay gisungog nga wala daw siyay amahan.< im_end > < im_start >user Tuga-tuga og langoy, hapit nuon nalumos!< im_end > < im_start >assistant None< im_end > < im_start >user {input_sentence}< im_end > < im_start >assistant </pre>
--

Figure H.1: End-to-End Few-Shot (E2E FS, 7-shots) prompt template used for the monolithic end-to-end baseline.

```

<|im_start|>system
You are a Cebuano language expert. The input sentence contains a figuratively used idiom.
Rewrite the entire sentence by replacing the idiomatic phrase with its literal Cebuano equivalent.

CRITICAL: Provide ONLY the direct answer. Do NOT include explanations, "Explanation:", or any other text.<|im_end|>
<|im_start|>user
Katagakon na akong mata, ugma nalang ni nato tiwason.<|im_end|>
<|im_start|>assistant
Katugon na ko, ugma nalang ni nato tiwason.<|im_end|>
<|im_start|>user
Kadugay jud ninyo nagpakasal? Lapas na man mo sa kalendaryo!<|im_end|>
<|im_start|>assistant
Kadugay jud ninyo nagpakasal? Lapas na man mo sa trayntay uno nga pangedaron!<|im_end|>
<|im_start|>user
Nihilak ang transferee kay gisungog nga nalumos daw siya sa sabaw.<|im_end|>
<|im_start|>assistant
Nihilak ang transferee kay gisungog nga wala daw siyay amahan.<|im_end|>
<|im_start|>user
{input_sentence}<|im_end|>
<|im_start|>assistant

```

Figure H.2: Few-Shot (FS, 3-shots) prompt template used for the Stage 2 normalizer baseline without boundary tags or dictionary hints.

```

<|im_start|>system
You are a Cebuano language expert. Rewrite the given sentence by removing the idiomatic phrase enclosed in <idiom></idiom> tags
and replacing it with a literal Cebuano equivalent.

Use any provided hint to understand the meaning, but feel free to modify root words, apply affixation, insert necessary function
words (like pronouns or enclitic particles), or use better synonyms so that the final sentence is grammatically perfect and
natural.

CRITICAL: Provide ONLY the direct answer. Do NOT include explanations, "Explanation:", or any other text.<|im_end|>
<|im_start|>user
Hints:
- The idiom "katagakon ang mata" generally means something similar to the following: [katugon, giduka]

<idiom>Katagakon na akong mata</idiom>, ugma nalang ni nato tiwason.<|im_end|>
<|im_start|>assistant
Katugon na ko, ugma nalang ni nato tiwason.<|im_end|>
<|im_start|>user
Hints:
- The idiom "lapas sa kalendaryo" generally means something similar to the following: [lapas sa trayntay uno nga pangedaron,
hamtong na]

Kadugay jud ninyo nagpakasal? <idiom>Lapas na man mo sa kalendaryo</idiom>!<|im_end|>
<|im_start|>assistant
Kadugay jud ninyo nagpakasal? Lapas na man mo sa trayntay uno nga pangedaron!<|im_end|>
<|im_start|>user
Hints:
- The idiom "nalumos sa sabaw" generally means something similar to the following: [walay amahan]

Nihilak ang transferee kay gisungog nga <idiom>nalumos daw siya sa sabaw</idiom>.<|im_end|>
<|im_start|>assistant
Nihilak ang transferee kay gisungog nga wala daw siyay amahan.<|im_end|>
<|im_start|>user
Hints:
{hints}

{tagged_input_sentence}<|im_end|>
<|im_start|>assistant

```

Figure H.3: Few-Shot with hints and tags (FS + hints + tags) prompt template used for the retrieval-grounded few-shot Stage 2 normalizer.

```

<|im_start|>system
You are a Cebuano language expert. The input sentence contains a figuratively used idiom.
Rewrite the entire sentence by replacing the idiomatic phrase with its literal Cebuano equivalent.

CRITICAL: Provide ONLY the direct answer. Do NOT include explanations, "Explanation:", or any other text.<|im_end|>
<|im_start|>user
{input_sentence}<|im_end|>
<|im_start|>assistant
{target_paraphrase}

```

Figure H.4: Fine-Tuned (FT) prompt template used for training the Stage 2 normalizer directly on raw sentences.

```
<|im_start|>system
You are a Cebuano language expert. The input sentence contains a figuratively used idiom. Rewrite the entire sentence by replacing
the idiomatic phrase with its literal Cebuano equivalent.

Hints:
{hints}

Use any provided hint to understand the meaning, but feel free to modify root words, apply affixation, insert necessary function
words (like pronouns or enclitic particles), or use better synonyms so that the final sentence is grammatically perfect and
natural.

CRITICAL: Provide ONLY the direct answer. Do NOT include explanations, "Explanation:", or any other text.<|im_end|>
<|im_start|>user
{input_sentence}<|im_end|>
<|im_start|>assistant
{target_paraphrase}
```

Figure H.5: Fine-Tuned with hints (FT + hints) prompt template used for training our optimal Stage 2 normalizer directly on raw sentences.

You are an expert evaluator for Cebuano sentence normalization.

Task:
Evaluate multiple candidate normalizations against a reference (gold) normalization.

Definition:
Sentence normalization means correctly replacing an idiomatic expression with its intended literal meaning while preserving the rest of the sentence.

Input:
 * Original Sentence: {sentence_with_idiom}
 * Reference Normalization: {gold_sentence}
 * Candidate Outputs:
 {candidate_list_text}

Evaluation Criteria (score each from 1 to 5):

1. Semantic Fidelity:
Does the candidate preserve the same overall meaning as the reference normalization?

Scoring guide:
 5 = Meaning fully preserves the intended meaning of the reference normalization
 4 = Minor differences, same meaning
 3 = Partially correct meaning
 2 = Weak or incomplete meaning overlap
 1 = Completely incorrect or unrelated meaning

CRITICAL RULES:
 - If the idiom is NOT normalized (e.g., copied verbatim, left unchanged, or only wrapped in tags such as <idiom></idiom>), semantic fidelity MUST be 1.
 - A sentence that copies the original without replacing the idiom has NOT conveyed the intended meaning.

2. Span Accuracy:
Does the candidate correctly replace ONLY the idiomatic expression and leave all other parts unchanged?

Scoring guide:
 5 = Idiom correctly replaced with minimal necessary changes to the rest of the sentence
 4 = Correct replacement with very minor unnecessary edits
 2 = Incorrect replacement or unnecessary modifications to other parts
 1 = Idiom not replaced OR output is structurally incorrect

CRITICAL RULES:
 - If the idiom is copied, left unchanged, or only wrapped in tags (e.g., <idiom></idiom>), the score MUST be 1.
 - Correct identification of the idiom WITHOUT replacing it is NOT sufficient.
 - Full sentence rewrites or loss of original structure MUST be penalized.

3. Fluency:
Does the candidate produce a grammatically natural, coherent, and well-formed Cebuano sentence?

Evaluate fluency independently of semantic fidelity and span accuracy. Focus only on linguistic quality, including grammatical correctness, readability, coherence, and natural sentence flow.

Scoring guide:
 5 = Fully natural, grammatically fluent, and native-like
 4 = Minor awkwardness or slight stylistic inconsistency
 3 = Understandable but noticeably unnatural or disfluent
 2 = Major grammatical issues, severe awkwardness, or poor sentence coherence
 1 = Incoherent, severely malformed, or unreadable sentence

CRITICAL RULES:
 - Fluency evaluates linguistic quality only, NOT correctness of meaning, idiom normalization accuracy, or similarity to the reference normalization.
 - A semantically incorrect candidate may still receive a high fluency score if it is grammatically natural and coherent.
 - Outputs containing nonsensical, non-existent, or contextually implausible lexical items MUST NOT receive a fluency score above 2.
 - Do NOT penalize grammatical imperfections already present in the reference normalization.
 - Outputs containing annotation tags (e.g., <idiom></idiom>) or formatting artifacts MUST NOT receive a fluency score above 3.
 - When uncertain between two adjacent scores, prefer the lower score unless the candidate clearly satisfies the higher category.

Instructions:
 * Evaluate EACH candidate independently.
 * CRITICAL: You MUST provide exactly {n_candidates} evaluations.
 * The number of objects in the "evaluations" array MUST be exactly {n_candidates}.
 * Do NOT skip any candidate. Do NOT combine multiple candidates into one evaluation.
 * Be strict and consistent across all candidates.
 * Use the full scoring range when appropriate; do not avoid low scores.
 * Do NOT reward partial solutions that fail to normalize the idiom.
 * Output STRICT JSON format only.

Figure I.1: Prompt template and grading rubric used for the LLM-as-a-judge evaluation of normalization candidates (continued on next page).

```
CRITICAL JSON FORMATTING RULES (failure to follow these will make the output unparseable):
* ALL string values - including "reasoning" and "summary_justification" - MUST be enclosed in double quotation marks ("").
* Do NOT write bare unquoted text as a JSON value.
* Numeric fields (candidate_id, semantic_fidelity, span_accuracy, fluency) must NOT be quoted.
* Do not add trailing commas.
* Do not include any text outside the JSON block.

Output STRICT JSON format:
{
  "evaluations": [
    {
      "candidate_id": <int>,
      "semantic_fidelity": <1-5>,
      "span_accuracy": <1-5>,
      "fluency": <1-5>,
      "reasoning": "<MUST be a quoted string - one sentence explaining the scores for this specific candidate>"
    }
  ],
  "summary_justification": "<MUST be a quoted string - concise pattern-level explanation of common errors across all candidates>"
}
```

Figure I.1: Prompt template and grading rubric used for the LLM-as-a-judge evaluation of normalization candidates (continued).

Error	Idiom (Meaning)	Input Sentence & Literal English Meaning	Config	Baseline Output & English Meaning	Expected Output & Figurative English Meaning
Syntax hallucination	<i>abli ang gaisano</i> (fly/zipper is open)	Piskot, wala ko kabantay nga abli akong Gaisano. (<i>Shoot, I did not notice that my Gaisano was open.</i>)	Zero-shot	Piskot, wala ko kabantay nga mag-buy akong Gaisano. (<i>Shoot, I did not notice that my Gaisano would-buy.</i>)	Piskot, wala ko kabantay nga abli akong zipper. (<i>Shoot, I did not notice that my zipper was open.</i>)
Literal amplification/Tagalog language switching	<i>abot sa pikas bukid</i> (someone who talks loudly)	Imong tingog kay abot sa pikas bukid. (<i>Your voice reaches the other mountain.</i>)	Zero-shot	Imong tingog kay abot sa totoong pikas bukid. (<i>Your voice reaches the real [in Tagalog] other mountain.</i>)	Imong tingog kay kusog kaayo. (<i>Your voice is very loud.</i>)
Literal interpretation	<i>abli ang gaisano</i> (fly/zipper is open)	Palihog ko pagsara, abli ang Gaisano. (<i>Please close it, the zipper is open.</i>)	Few-shot	Palihog ko pagsara, abli ang Gaisano shopping center. (<i>Please close it, the Gaisano shopping center is open.</i>)	Palihog ko pagsara, abli ang zipper. (<i>Please close it, the zipper is open.</i>)
Verbatim copying	<i>abot na sa dukot</i> (running out of resources)	Nagsige na lang mi og gamit sa daan nga mga gamit kay abot nas dukot ang suplay sa ospital. (<i>We just keep on using the old equipment because the hospital supplies have reached the burnt rice.</i>)	Few-shot	Nagsige na lang mi og gamit sa daan nga mga gamit kay abot nas dukot ang suplay sa ospital. (<i>We just keep on using the old equipment because the hospital supplies have reached the burnt rice.</i>)	Nagsige na lang mi og gamit sa daan nga mga gamit kay pahurot na ang suplay sa ospital. (<i>We just keep on using the old equipment because the hospital supplies are running out.</i>)
Unknown word insertion	<i>ander</i> (henpecked/undercontrol)	Ayaw pa-ander sa imong uyab, pre. (<i>Do not let yourself be under your girlfriend, friend.</i>)	Few-shot	Ayaw pa-ander sa imong uyab, panatangani. (<i>Do not be under your girlfriend's control, panatangani (unknown word).</i>)	Ayaw pa-kontrol sa imong uyab, pre. (<i>Do not let yourself be controlled by your girlfriend, friend.</i>)
In-context leakage from one of Few-Shot samples in template	<i>abot na sa dukot</i> (running out of resources)	Nagsige na lang mi og gamit sa daan nga mga gamit kay abot nas dukot ang suplay sa ospital. (<i>We just keep on using the old equipment because the hospital supplies have run out.</i>)	Few-shot	Nagsige na lang mi og gamit sa daan nga mga gamit kay walay amahan ang suplay sa ospital. (<i>We keep using old equipment because the hospital supplies have no father.</i>)	Nagsige na lang mi og gamit sa daan nga mga gamit kay pahurot na ang suplay sa ospital. (<i>We just keep on using the old equipment because the hospital supplies are running out.</i>)

Table J.1: Examples of baseline normalization errors under untuned settings without dictionary hints.

Idiom (Meaning)	Input Sentence & Literal English Meaning	Config (Stage 1 and Stage 2)	Baseline Output	Expected Output & Figurative English Meaning
<i>nabiyaan sa tren</i> (left behind/failed to adapt)	Sa <idiom>papel</idiom> gihapon ka magdrawing? Nabiyaan na ka sa tren, naa nay digital art karon. (Are you still drawing on paper? You were left behind the train, there is digital art now.) [Imperfectly tagged by Stage 1]	5-tag BIO and FS + hints + tags	Sa <idiom>papel</idiom> gihapon ka magdrawing? Wala kaapas ka magbag-o, naa nay digital art karon. (Are you still drawing on <idiom>paper</idiom>? You were left behind, there is digital art now.) [Broken Cebuano Grammar]	Sa papel gihapon ka magdrawing? Wala na ka kaapas sa bag-o, naa nay digital art karon. (Are you still drawing on paper? You were left behind, there is digital art now.)
<i>agatha</i> (stubborn/closed-minded)	Lisod ikasabot ang tawo nga <idiom>Agatha</idiom> kay sarado ang utok sa laing suhestiyon. (It is hard to agree with a person like Agatha because their mind is closed to other suggestions.)	5-tag BIO+TL+CRF and FT + tags	Lisod ikasabot ang tawo nga matinu-manon sa gusto kay sarado ang <idiom>utok</idiom> sa laing suhestiyon. (It is hard to agree with a stubborn person because their mind is closed to other suggestions.)	Lisod ikasabot ang tawo nga matinu-manon sa gusto kay sarado ang utok sa laing suhestiyon. (It is hard to agree with a stubborn person because their mind is closed to other suggestions.)
<i>abli ang gaisano</i> (fly/zipper is open)	Palihog ko pagsara, <idiom>abli</idiom> ang Gaisano. (Please close it, the Gaisano is open.) [Truncated tagging by Stage 1]	5-tag BIO+TL and FT + tags	Palihog ko pagsara, gastador ang Gaisano. (Please close it, Gaisano is a spendthrift.)	Palihog ko pagsara, abli ang zipper. (Please close it, the zipper is open.)

Table K.1: Examples of XML tag leakage, tag shifting, and boundary truncation under tagged configurations.

Metric	Complete	No-Literal	No-HN
Detection F1 (%)	87.10 $\pm$ 1.09	98.71 $\pm$ 0.55	92.35 $\pm$ 2.37
Strict Match (%)	90.35 $\pm$ 0.85	98.68 $\pm$ 0.57	90.83 $\pm$ 1.44
BLEU	86.80 $\pm$ 0.81	96.22 $\pm$ 1.04	84.63 $\pm$ 2.60
Composite (1–5)	4.61 $\pm$ 0.00	4.98 $\pm$ 0.03	4.68 $\pm$ 0.11

Table L.1: Performance of the CebIDN pipeline (utilizing CebBERT 3-tag BIO+CRF and Qwen-2.5-7B) across Complete, No-Literal, and No-HN idiom ambiguity classes, averaged over three evaluation seeds.

Configuration	Stage 2 Input & Literal English Meaning	IKB Match & Hint (with English Translation)	Stage 2 Output & Figurative English Meaning
3-tag BIO+CRF (Tagged FT+hints)	<idiom>Abot sa dalunggan ang ngisi</idiom> sa bata sa dihang gihatagan siya’g regalo. (The child’s smile reached his/her ears when he/she was given a gift.)	Match: abot sa dalunggan ang ngisi (<i>smile reached the ears</i>) Hint: hilabihan ang kalipay (<i>extremely happy</i>)	Hilabihan ang kalipay sa bata sa dihang gihatagan siya’g regalo. (The child was extremely happy when they were given a gift.)
3-tag BIO+CRF (Tagged FT+hints)	Si Pedro kay <idiom>kandingon kaayo</idiom> na sukad pa pagkabata. (Pedro has been very goat-like since childhood.)	Match: kandingon (<i>goat-like</i>) Hint: tapulan maligo (<i>lazy to take a bath</i>)	Si Pedro kay tapulan kaayo maligo na sukad pa pagkabata. (Pedro has been very lazy to take a bath since childhood.)
3-tag BIO+CRF (Tagged FT fallback)	Palihog ko pagsara, <idiom>abli</idiom> ang Gaisano. (Please close it, Gaisano is open.)	Match: None Hint: None	Palihog ko pagsara, abli ang zipper. (Please close it, the zipper is open.)
3-tag BIO (Tag-Free FT+hints)	Pag-weekend, kandingon g’yud kaayo nang bayhana, kutob ra g’yud intawon sa hilaam-os. (On weekend, that woman is really goat-like, she only goes as far as washing her face.)	Match: kandingon (<i>goat-like</i>) Hint: tapulan maligo (<i>lazy to take a bath</i>)	Pag-weekend, tapulan g’yud kaayo nang bayhana maligo, kutob ra g’yud intawon sa hilaam-os. (On weekend, that woman is really lazy to take a bath, she only goes as far as washing her face.)
3-tag BIO (Tag-Free FT fallback)	Piskot, wala ko kabantay nga abli akong Gaisano. (Shoot, I didn’t notice my Gaisano was open.)	Match: None Hint: None	Piskot, wala ko kabantay nga abli akong zipper. (Shoot, I didn’t notice my zipper was open.)

Table M.1: Qualitative success cases showing the normalization workflow across different configurations. The input to Stage 2 (column 2) contains boundary tags in the tagged normalizer configuration (Cases 1–3) but is tag-free in the tag-free normalizer configuration (Cases 4–5). Columns 3 and 4 show the retrieved Idiomatic Knowledge Base (IKB) hints and final output sentences, respectively.