Not Just a Piece of Cake: Cross-Lingual Fine-Tuning for Idiom Identification

Ofri Hefetz and **Kai Golan Hashiloni** and **Alon Mannor** and **Kfir Bar**
Efi Arazi School of Computer Science, Reichman University, Herzilya, Israel
{ofri.hefetz, kai.golanhashiloni, alon.mannor}@post.runi.ac.il,
kfir.bar@runi.ac.il

Abstract

We investigate cross-lingual fine-tuning for idiomatic expression identification, addressing the limited availability of annotated data in many languages. We evaluate encoder and generative decoder models to examine their ability to generalize idiom identification across languages. Additionally, we conduct an explainability study using linear probing and LogitLens to analyze how idiomatic meaning is represented across model layers. Results show consistent cross-lingual transfer, with English emerging as a strong source language. All code and models are released to support future research.

1 Introduction

Idiomatic expressions are multiword constructions whose meaning cannot be deduced from the meanings of their individual words. They are typically fixed or semi-fixed in form and exhibit semantic non-compositionality, where the overall meaning differs from the literal meanings of the parts. As a subclass of multiword expressions (MWEs), idioms present a significant challenge for computational models due to their ambiguity, variability, and cultural grounding.

In this paper, we focus on potential idiomatic expressions (PIEs), which are phrases such as *spill the beans* or *break the ice* that may appear either idiomatically or literally, depending on context. The goal of the task is not merely to recognize idioms in general, but to identify all PIEs that are used idiomatically in context within a given text. This form of contextual disambiguation is essential: *spill the beans* in *I spilled the beans during the meeting* conveys a figurative meaning, whereas the same phrase in *I spilled the beans in the kitchen* does not.

Identifying idiomatic usage is particularly challenging in multilingual and low-resource contexts. Most previous work adopt a monolingual approach and depends on annotated datasets, which are often scarce and expensive to create, especially for low-resource or historical languages, where idioms play a crucial role in interpretation.

To address these limitations, we explore cross-lingual transfer, which involves using annotated data in one language to improve idiom identification in another. We hypothesize that large multilingual Large language models (LLMs) already encode latent cross-lingual representations that can be fine-tuned for this task. This follows findings such as AlignXIE (Zuo et al., 2024), which showed multilingual alignment improves zero-shot Named Entity Recognition (NER) and inspired us to examine whether similar transfer can handle the more complex phenomenon of idiomaticity.

To this end, we present the first systematic study of cross-lingual fine-tuning (FT) for idiomatic expression identification. We evaluate both encoder-based and generative decoder models, fine-tuning them on source-language data and assess their performance on typologically distinct target languages. In addition to standard evaluation, we probe the models' internal representations to understand how non-compositional meaning is encoded and transferred across languages.

Our main contributions are as follows: **(1)** We introduce the first systematic study of cross-lingual transfer learning for idiomatic expression identification using language models, based on supervised fine-tuning. **(2)** We evaluate the potential of using annotated data from one language to improve idiom identification in target languages, offering insight into the feasibility of such transfer. **(3)** We perform a structural and representational analysis of idiomatic expressions, probing how models capture non-compositional semantics across languages.

To guide our investigation, we pose the following research questions: **RQ1:** Can LLMs, when fine-tuned on idiom identification in one language,

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2521–2537
December 20-24, 2025 ©2025 Association for Computational Linguistics

generalize to accurately identify idiomatic expressions in other languages? **RQ2:** Does fine-tuning on a mixture of languages improve a model's ability to recognize idiomatic expressions? **RQ3:** Do the internal representations of LLMs differentiate between idiomatic and literal usages?

We release all models, code, prompts, and evaluation scripts to ensure reproducibility.[1]

2 Related Work

2.1 Idiom Identification

Idiomatic expressions, a subclass of MWEs, have non-compositional meanings that cannot be inferred from their parts (Baldwin and Kim). Detecting idioms requires identifying their span in text and distinguishing literal from figurative use. Early approaches relied on syntactic and statistical cues (Fazly et al., 2009; Cook et al., 2007; Shutova et al., 2010), later enhanced by semantic vector representations (Gharbieh et al., 2016; Nedumpozhimana and Kelleher, 2021). More recent work has applied neural models, including CNNs, RNNs, and transformers, to idiom classification (Zeng and Bhat, 2021; Briskilal and Subalalitha, 2022; He et al., 2024), typically under the assumption that the PIE has already been identified and the task is to classify it as idiomatic or literal.

Following recent span-based idiom and MWE identification work in monolingual settings (Hashiloni et al., 2025), we require models to identify PIEs in context and determine whether they are used figuratively, without access to the spans in advance. We extend this formulation to the cross-lingual setting and evaluate both encoder- and decoder-based models under multilingual transfer. This makes the task especially challenging in multilingual and low-resource settings, where idiom inventories, cultural grounding, and figurative usage patterns vary substantially across languages.

2.2 Multilingual Idiom Identification Datasets

Idioms are well-studied in English, but multilingual idiomaticity has only recently gained attention. PARSEME (Ramisch et al., 2020) covers verbal MWEs in 20+ languages, offering a basis for cross-lingual research, though its broad definition of verbal idioms limits its use for idiom identification.

English datasets such as VNC-Tokens (Cook et al., 2008), SemEval-2013 Task 5 (Korkontzelos et al., 2013), and IDIX (Sporleder et al., 2010) are foundational but small, language-limited, and cover few idiom types (Mi et al., 2025; Arslan et al., 2025; Haagsma et al., 2020; Tedeschi et al., 2022). More recent sets like AStitchInLanguageModels (Tayyar Madabushi et al., 2021), EPIe (Saxena and Paul, 2020), and SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022) expand coverage but remain mostly English-centric and focus on classification. We exclude them due to overlap with our evaluation datasets. For multilingual evaluation, we use four idiom-specific datasets: Dodiom (Eryiğit et al., 2022), ID10M (Tedeschi et al., 2022), MAGPIE (Haagsma et al., 2020), and the OpenMWE Corpus (Hashimoto and Kawahara, 2009) - which together cover eleven languages. Section 3.2 details their annotations, task formats, and processing.

Note that while prior multilingual resources such as PARSEME and the SemEval shared tasks have been instrumental in advancing idiom research, their goals differ from ours. Our combined corpus focuses specifically on idiomatic expressions exhibiting clear non-compositional meaning, offering both a larger scale and higher idiom diversity.

2.3 Fine-Tuning LLMs

Encoder models. Encoder models, such as BERT (Devlin et al., 2019; Lee and Hsiang, 2020) and XLM-RoBERTa (Conneau et al., 2020), are standard for structured prediction. Idiom identification typically uses them for binary classification (literal vs. idiomatic) (Chakrabarty et al., 2021; Briskilal and Subalalitha, 2022) and span identification (Tedeschi et al., 2022). Their success in cross-lingual NER and transfer learning (Pfeiffer et al., 2021; Parovic et al., 2023) makes them strong baselines for idiom identification.

Decoder models. Recently, decoder-based models like GPT (Li et al., 2024) and Llama (Šmíd et al., 2024) have been fine-tuned for open-ended tasks, showing strong performance in span extraction and structured generation, making them promising for multilingual idiom identification. Ide et al. (2025) introduced the CoAM dataset for MWE identification, where large generative models such as Qwen-2.5-72B outperformed prior methods, highlighting a shift toward decoder-based approaches. Arslan et al. (2025) extended this to

[1] https://github.com/Intellexus-DSI/Not-just-a-piece-of-cake

multilingual idioms, creating synthetic corpora in four languages and fine-tuning both encoder and decoder models. Despite these advances, idiom identification, especially in cross-lingual settings, remains underexplored, and our work systematically investigates cross-lingual fine-tuning and evaluation across a broader set of natural multilingual idiom datasets.

3 Method

3.1 Problem Formulation

We treat idiom identification as a token classification task, following Tedeschi et al. (2022) and Ide et al. (2025). For a given sentence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ where each $\mathbf{x_i}$ is a single token. The goal is to assign a sequence of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$, where each $y_i \in \{\texttt{B- IDIOM}, \texttt{I-IDIOM}, \texttt{O}\}$ indicates whether the token begins, continues, or is outside an idiomatic expression. This annotation scheme is commonly known as BIO tagging.

This task encompasses both identification and disambiguation elements, requiring models to not only identify idiomatic expressions but also distinguish between figurative and literal usage. Notably, no candidate expressions are provided in advance, enhancing the realism of the task.

To address this, we fine-tune both encoder-based models (e.g., XLM-RoBERTa, BERT) for token classification and decoder-based models (e.g., Llama) for generative prediction of idiom spans. In both cases, the final output is evaluated using BIO-formatted labels aligned with the gold labels. Because evaluation is performed at the token level, both boundary mismatches (e.g., predicting a longer or shorter span) and errors in distinguishing figurative from literal usage are directly reflected in the resulting F1 scores.

3.2 Datasets

We utilize four datasets to capture the multilingual and structural diversity of idioms, adhering to their original licenses and purposes. Table 1 shows statistics, with license details in Appendix I. Our experiments include training data from eleven languages: English (EN), German (DE), Italian (IT), Spanish (ES), Turkish (TR), Japanese (JA), Dutch (NL), French (FR), Dutch (NL), Chinese (ZH), and Portuguese (PT). Evaluation is carried out on six languages - EN, DE, IT, ES, TR, and JA for which test data is either publicly available or derived by partitioning the dataset when official splits are not available.

Dodiom (Eryiğit et al., 2022). Dodiom is a multilingual dataset for idiom identification, constructed via a gamified crowd-sourcing approach targeting Turkish and Italian. Idioms were drawn from existing resources, including PARSEME 1.3 (Savary et al., 2023) and online lists, with a subset manually validated by expert linguists to create a gold standard. Lacking a predefined split, we created our own and converted spans to token-level BIO tags for evaluation (details in Appendix A).

ID10M (Tedeschi et al., 2022). A multilingual idiom identification dataset with *silver-standard* training data in ten languages, built from Wiktionary,[2] idioms contextualized via WikiMatrix. The test set includes 200 gold examples in English, German, and Italian, and 199 in Spanish (despite the paper stating 200). It contains only continuous idiom spans, where all idiom words appear consecutively without gaps or intervening tokens, supports multiple idioms per instance, and uses BIO tagging.

MAGPIE (Haagsma et al., 2020). MAGPIE is a gold-standard English dataset of 56,622 examples from diverse genres. Initially built for idiom classification, we adapt it for token-level classification (Appendix B). After preprocessing, the test split consists of 4,391 samples, from which we evaluate 400 due to inference budget limits. This version is marked as *ours* in Table 1.

OpenMWE Corpus (Hashimoto and Kawahara, 2009). The OpenMWE Corpus is a large Japanese dataset for classifying idioms as figurative (`I`) or literal (`L`), containing 102,334 examples for 146 idioms from a web corpus. We convert its annotations to token-level BIO format and create our own train/test split, as none is publicly available. Full preprocessing details are in Appendix C.

3.3 Fine-Tuning

To study idiom identification in multilingual settings, we fine-tune two language model types: (1) encoder models using token-level classification, and (2) instruction-tuned LLMs with supervised fine-tuning. All models considered in this work were originally pre-trained on the target languages

[2] https://www.wiktionary.org/

Dataset	**Language**	**Train**		**Test**	
		# Sentences	**# PIE**	**# Sentences**	**# PIE**
Id10M	EN	37,919	4,568	200	142
	DE	24,126	819	200	111
	ES	25,070	1,229	199	78
	IT	26,372	452	200	139
	JA	6,388	165	-	-
	NL	18,801	189	-	-
	FR	31,807	188	-	-
	PL	31,963	648	-	-
	PT	27,594	559	-	-
	ZH	8,249	1,301	-	-
OPENMWE	JA	100,503	146	577	146
MAGPIE	EN	35,542	27,296	4,451	3,401
MAGPIE (our)	EN	34,136	26,907	400	298
Dodiom	TR	6,361	36	500	36
	IT	7,033	38	500	38

Table 1: Dataset statistics used in our experiments.

to avoid zero-resource scenarios. Checkpoints, parameter sizes, and licenses are detailed in Appendices J and I.

3.3.1 Fine-Tuning Decoder Models

We fine-tune two instruction-tuned Llama models[3] (Grattafiori et al., 2024) for our experiments: `Meta-Llama-3.1-8B-Instruct-Reference` and `Meta-Llama-3.1-70B-Instruct-Reference`, both accessed via the Together AI platform.[4]

To tailor the models to our idiom identification task, we explore several prompt formats during development (see Appendix F). We ultimately converge on a structured schema consisting of a *system prompt*, a *user prompt*, and an *assistant response*, which yields the most accurate and format-consistent outputs (see Figure 9 located in Appendix F).

Given the scale of the models, we adopt LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. The complete set of training hyperparameters is provided in Table 2, located in Appendix D.1.

To control compute costs while preserving linguistic diversity, we randomly sample 1,500 training examples per language. To evaluate the robustness of our results, we repeated this sampling process using three different random seeds. Each subset was then used to fine-tune a separate instance of the Llama 8B model. In contrast, the 70B model is fine-tuned once due to resource constraints.

Each model is fine-tuned separately for each source language (excluding CZ and JA) and evaluated across six target languages.

Cross-lingual transfer with language mixtures: In addition to the standard language-to-language transfer fine-tuning, we explore two other configurations in which we included multiple languages in the training set: (i) a small mixture that includes the six languages used for evaluation (EN (ID10M), DE, IT (ID10M), ES, TR, JA), (ii) a large mixture that also covers NL, FR, PL, and PT. All configurations use balanced sampling (1,500 examples per language). Each model was fine-tuned once, and inference was repeated three times with different seeds to assess variability.

Expected output format and evaluation. The decoder is instructed to return a minimal JSON object containing only the predicted idioms for the input sentence (no explanations), which we then convert into token-level BIO tags for scoring. Following Appendix E, we compute macro-averaged token-level F1. To align predicted idioms with token spans, we normalize both the sentence and predictions (e.g., lowercasing, quote/dash standardization) and apply character-level substring matching. Because alignment is non-trivial and minor formatting deviations can lead to drops in measured performance, the evaluation pipeline includes robust recovery for malformed outputs via relaxed JSON parsing. Full alignment and parsing details appear in Appendix E.

3.3.2 Fine-Tuning Encoder-Based Models

We also fine-tune encoder-based models using the full training sets, and each training and evaluation process repeats five times with different seeds to ensure robustness. We train multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) as multilingual models, both pre-trained on all target languages.

We exclude the Italian and Turkish from the Dodiom dataset, as well as the Chinese, Portuguese, and Dutch from ID10M as source training languages, as a preliminary experiment showed that the datasets are too small to reach any meaningful learning. For the list of hyperparameters used during training, see Appendix D.

We fine-tune multiple encoder-based models for idiom identification. While Tedeschi et al. (2022) followed a comparable approach, we adopt a minimal fine-tuning setup using the *HuggingFace* training framework (Wolf et al., 2020). The hyperparameters are provided in Appendix D.3.

Cross-lingual transfer with language mixtures. We further investigate cross-lingual transfer learning in encoder-based models through fine-

[3] `https://ai.meta.com/blog/meta-llama-3-1/`

[4] `https://www.together.ai/`

tuning experiments. The model is trained on data from **all eleven languages** included in our study, while evaluation is conducted on a **five-language test set** comprising English (EN), German (DE), Italian (IT), and Spanish (ES) from the ID10M dataset, and Japanese (JA) from the OPENMWE dataset. To explore the impact of cross-lingual transfer, we fine-tune mBERT using training data composed of $X\%$ from the target language and $(100 - X)\%$ from a balanced mix of the remaining languages, where $X \in \{100, 90, \ldots, 0\}$. To allow for a balanced sampling from all languages, and since Turkish has the minimum number of 6,361 samples in its training set, we cap the Japanese at 60,000 examples (for $X = 0$, we include 6K samples of each language). We evaluate performance on the test set of the target language in each case. To ensure fairness, we keep the total training size equal to the original size of the target language's training set.

To ensure robustness, we repeat each experiment across five random seeds, where the seed influences data sampling, shuffling, and training.

Hyperparameter selection. We tune all transformer baselines on English ID10M and fix the best per-model hyperparameter configuration for cross-lingual runs to ensure comparability; full search spaces, and outcomes are in Appendix D.3.

3.4 Explainability Analysis

To examine how LLMs encode idiomatic expressions, we conduct a layer-wise probing analysis (Conneau et al., 2018; Hewitt and Manning, 2019), treating each Transformer layer as a representation state. For each dataset sentence, we pass it through a decoder model, extracting hidden states from every layer. We compute an element-wise average over only the tokens annotated as part of the PIE. For example, for *break the ice*, only *break* and *ice* are included, as *the* is unannotated, a standard convention. This produces one vector per expression per layer, which we use to compare idiomatic and literal instances across contexts. This analysis relies on precise annotations available in MAGPIE (EN) and Dodiom (TR, IT). We obtain these representations from Llama-3.2-3B to examine how idiomatic and literal usages are encoded across the model's layers. Our investigation includes three types of analysis:

Linear probing using Logistic Regression. For each layer, we train a logistic regression classifier on PIE representations (produced as described above) with *idiomatic* vs. *literal* labels. We sampled 1,000 training examples from each dataset per language. For IT and TR, the datasets contain relatively few unique PIE (see Table 1), so we sampled without restrictions, resulting in 1,000 examples covering 33 PIEs in IT and 36 in TR. For English, which has a much larger idiom inventory, we restricted sampling to 50 unique PIE types to match the cross-lingual pattern and drew 1,000 examples accordingly.

We then train a logistic regression classifier on the training-set vectors using `scikit-learn`'s implementation (Pedregosa et al., 2011) with `max_iter=1,000` and report F1 on the test-set vectors to assess how linearly separable idiomaticity is across layers. MAGPIE is evaluated on its full filtered test set (4,278 samples), while for Dodiom, we create a new test set of the same size.

Multilingual linear probing. We train a logistic regression classifier on each language and evaluate its performance on the test set of the other two languages. This helps us to evaluate the transfer-learning capability of the model through the information encoded in its hidden states. We repeat this process for each layer as before.

`LogitLens` probing. We use LogitLens probing (Belrose et al., 2023) to examine whether idiomatic and literal usages yield similar token-level predictions across layers. LogitLens projects hidden states through the language modeling head, showing each layer's "intended" token distributions and providing a layer-wise view of semantic evolution.

From each dataset, we sample 5,000 instances containing a single PIE, balancing idiomatic and literal usages by retaining equal counts per idiom. This yields 192 English idioms (avg. 4 instances), 29 Italian (avg. 129), and 36 Turkish (avg. 117), reflecting dataset-specific annotation methods.

For each layer, we apply the model's normalization and LM head to compute intermediate logits and extract the top-5 predicted tokens, representing each occurrence as a token set. We then compute per-layer Jaccard similarity between idiomatic and literal sets. Averaging across PIEs produces a similarity curve: lower scores indicate greater divergence and stronger semantic separation between literal and idiomatic meanings.

4 Results

Details on computational cost and resources are provided in Appendix G.

4.1 RQ1: Cross-Lingual Generalization

Encoder results. To evaluate the cross-lingual generalization ability of encoder-based models for idiom identification, we fine-tuned two multilingual encoders: mBERT and XLM-RoBERTa on individual source languages and evaluated them across a variety of target languages. The results are visualized in Figures 1 and 8 in Appendix H.1, where each cell represents the average F1 score across five seeds. A redder/lighter color indicates higher performance, while a darker blue reflects lower scores. To aid interpretation, the heatmaps also display the rounded F1 scores (without decimal points) within each cell. For the exact numerical results, including standard deviations, refer to Table 6 in Appendix H.2.

The results reveal moderate but noteworthy cross-lingual transfer. For instance, mBERT trained on high-resource languages (top left corner on the heat map) for example: EN (ID10M) achieves F1 scores of 56.3 on DE, 61.8 on IT (Dodiom), and 59.1 on ES. This suggests that the model captures the concept of idiomacity and has a substantial ability to generalize it beyond the language it was trained on. Similarly, XLM-RoBERTa trained on EN (MAGPIE) reaches 49.3 on DE, 57.3 on IT (ID10M), and 40.0 on ES. While the most substantial transfer effects occur between high-resource languages, we also observe cross-family generalization (e.g., EN to JA or JA to TR), highlighting the ability of multilingual encoders to abstract figurative patterns across different language families. Although performance remains below in-language fine-tuning (e.g., mBERT EN→EN: 75.3).

Generative decoder results. We additionally examine generative decoder models. Figure 2 presents the difference in F1 score between various Llama 8B fine-tuning configurations and the zero-shot baseline, across different target languages (full results are provided in Table 8, Appendix H.3).

Overall, the results are mixed, with many setups showing slight drops, but several clear gains stand out. The top-right of the heatmap displays a cluster of light-colored cells, showing consistent improvements when training on high-resource languages and testing on low-resource ones (e.g., EN and DE boosting Japanese by +0.41 and +0.44). Dutch notably improves Turkish (+1.09) and Japanese (+0.42), Turkish boosts Italian-Dodiom (+0.58) and Spanish (+0.45), and Polish yields large gains on Spanish (+0.80).

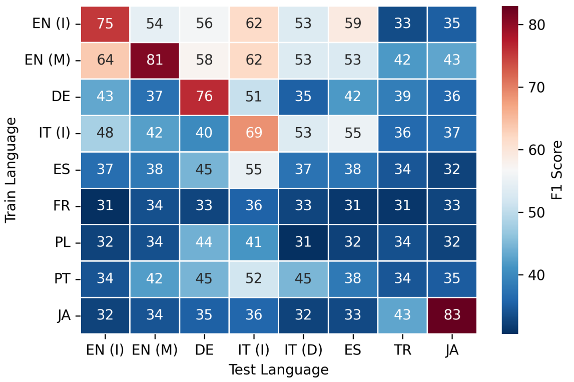

Figure 1: F1 scores of the FT mBERT. Cells show average macro F1 over five seeds. EN uses both ID10M (I) and MAGPIE (M) for training/eval; IT is trained on ID10M and evaluated on ID10M (I) and Dodiom (D).

These patterns suggest Llama 8B benefits from selective cross-lingual transfer, especially from high-to-low-resource or distant languages (e.g., EN→TR). However, some pairs harm performance, such as Italian training reducing German (−2.01) and English-ID10M (−0.80), highlighting the importance of source–target compatibility.

While most of our experiments focused on the Llama 8B model due to resource constraints, we fine-tuned the larger Llama 70B model once per configuration to assess the effect of model scaling. As shown in Table 8 in Appendix H.3, the 70B model did not consistently outperform its 8B counterpart. In fact, for many test languages, including EN, DE, and IT, the performance of the larger model decreased across most training configurations compared to the baseline. However, two notable exceptions emerged: ES and JA test languages. In these cases, the 70B model achieved consistently higher scores across nearly all training languages. For example, when testing on JA, performance increased by (+3.98) using PT for training, and ES showed similar gains (e.g., PL→ES: +1.31). These results suggest that model scaling may be particularly beneficial for languages with limited training data, such as JA. Nevertheless, the degradation in other languages suggests that the benefits of scale are not uniform and may interact with factors such as language similarity, resource availability, or training dynamics.

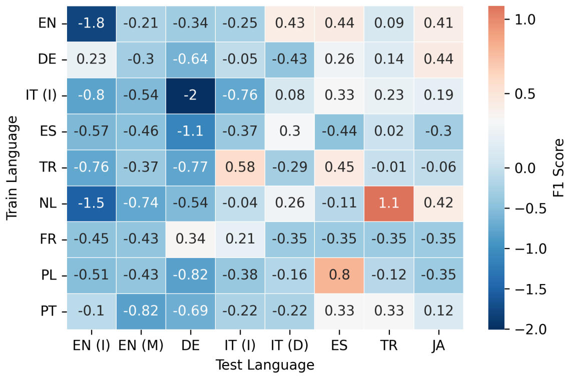

Figure 2: F1 differences between the non-fine-tuned Llama 8B baseline and its three fine-tuned variants. Each cell represents the average difference across three seeds. Dataset abbreviations are defined in Figure 1.

Summary (RQ1): Our results show that encoder and decoder LLMs can generalize idiom identification across languages from a single source language. Encoders exhibit moderate cross-lingual transfer, while decoders show more selective transfer, especially from high-to-low-resource languages, supporting the idea that idiomatic meaning encodes transferable semantic patterns.

4.2 RQ2: FT on Mixture of Languages

Encoder results. Figure 3 illustrates the effect of gradually increasing the proportion of target-language data in the training set during FT. which start from strong baselines and continue to benefit from additional in-language data. However, for these languages, even a small proportion (e.g., 10–20%) of target-language data yields substantial performance gains, highlighting the multilingual adaptability of encoder-based models like mBERT. In contrast, low-resource or typologically distant languages such as Japanese and Spanish start from lower zero-shot baselines and exhibit smaller overall gains. Spanish shows an unusual pattern: performance drops as more target-language data is added, unlike the other languages, where adding such data yields slight gains.

Examining the test sets suggests why. In Spanish, only 24.4% of the idioms in the test set also appear in the training set, compared to 43.7% for English, 63.9% for German, and 62.6% for Italian. At the same time, 84.6% of Spanish test instances are literal usages, the highest rate across all test sets (English: 28.9%, German: 17.1%, Italian: 34.5%). This means that most of the additional Spanish training data introduces idioms absent from the test set while also reducing the proportion of multilingual data in the mix. This combination appears harmful: although the added data matches the target language, it lacks overlap with the test idioms, and the reduced language diversity weakens cross-lingual generalization. These findings suggest that maintaining a balanced multilingual mix can be more effective than simply increasing target-language data, especially when the new examples do not align semantically with the evaluation set.

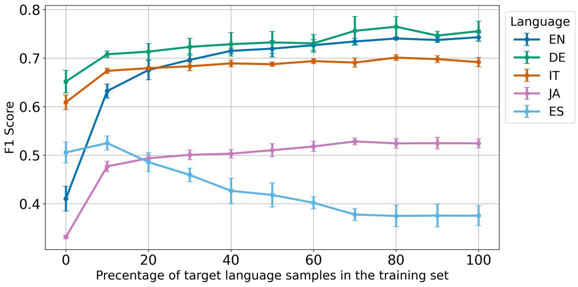

Figure 3: F1 scores of mBERT model fine-tuned on a mixed-language training set and evaluated on the target languages.

For Japanese, however, we could not identify a satisfying explanation for the observed performance pattern. The interplay between cultural specificity, idiom transparency, and typological distance remains unclear, and we plan to investigate this open question in future work.

Decoder results. To evaluate whether multilingual FT improves the ability of generative decoder models to identify idiomatic expressions, we compare the performance of Llama-3.1-8B under three training configurations: monolingual (baseline), a small multilingual mix (six languages), and a large multilingual mix (ten languages). Unlike our encoder experiments, we were unable to replicate the gradual target-language scaling due to resource constraints. The results are shown in Figure 4, with exact scores provided in Table 8 (Appendix H.3).

Overall, the small multilingual mix improves or matches baseline performance in most cases. For example, it improves performance on JA (from 42.38 to 43.43) and TR (from 46.32 to 46.59), which may stem from their lower baselines, leaving more room for improvement. In contrast, the large mix yields more variable results. While it maintains competitive performance on some targets (e.g., ES and IT (Dodiom)), it often underperforms compared to the small mix, particularly on DE and IT (ID10M), suggesting that increasing

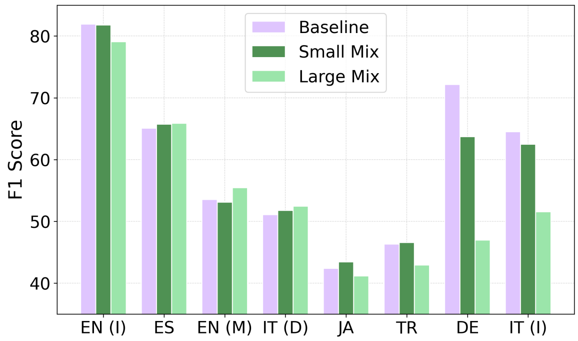

Figure 4: Average F1 scores of fine-tuned Llama 8B on a mixed-language training set and evaluated on individual target languages. Dataset abbreviations are defined in Figure 1.

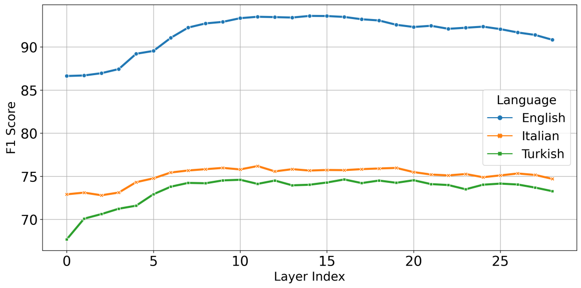

Figure 5: Llama-3.2-3B monolingual linear-probe F1 across layers.

the number of training languages may introduce interference or reduce focus on the target signal.

Summary (RQ2): Our findings show that in some cases, fine-tuning on a language mix improves idiom identification for both encoder- and decoder-based models. For encoders, even limited target-language data yields strong results in high-resource languages, making multilingual FT an efficient alternative to monolingual training. For decoders, a small multilingual mix often matches or outperforms baselines, especially for low-resource languages like Japanese and Turkish, though adding more languages produces mixed outcomes. Overall, multilingual FT aids cross-lingual generalization, but its impact depends on model type and training set composition.

4.3 RQ3: Representation of Idioms

Figure 5 shows **monolingual linear probing** results across layers. Performance rises until about layer 7, stabilizes through the mid-layers, and then slightly declines after layer 26. While the absolute differences between layers are modest, a consistent trend emerges: English maintains F1 scores close to 96 across layers 7–25, suggesting stable cues to non-compositionality. Italian and Turkish plateau just above and below 75, reflecting comparatively weaker but still steady representations. Overall, these results indicate that from layer 7 onward, decoder models capture idiomatic information stably, even if the layer-wise differences are not significant.

Figure 6 presents **Jaccard similarity** trends from LogitLens. All languages show a parabolic pattern with a minimum near layers 10–11, aligning with the linear probing results and indicating stronger idiomatic–literal separation from layer 7 upward. A qualitative analysis supports this observation. For the PIE *at the crossroads*, literal uses align with tokens like *crossings*, *intersection*, and *junction*, while idiomatic uses often evoke words such as *critical*, *risk*, and *resolving*, highlighting the shift from literal to figurative meaning. A rise in Jaccard similarity at the top layers (25 and onward) mirrors the slight F1 drop in linear probing in Figure 5, suggesting reduced sensitivity of LogitLens near the output without major impact on distinguishing idiomatic from literal usages.

The **cross-lingual linear probing** results are shown in Figure 7. For English as a source, we compare two training set compositions. The first, EN, follows Section 3.4 and uses a balanced sample of 50 PIE types to match the Italian and Turkish distributions. The second, EN2, removes this constraint and samples 1,000 examples uniformly from the full English dataset, resulting in 574 distinct PIEs. This setup tests whether English's larger idiom inventory improves transfer. Both configurations show similar trends when transferring to Italian and Turkish, with F1 scores around 70 and stable performance across layers, suggesting English is a strong source regardless of sample balance. TR performs poorly as a source, especially on IT (F1 < 10), while IT→TR transfers better, indicating asymmetry and typological divergence. Most settings show a drop in upper layers, except English in both configurations, which remains stable across targets.

Summary (RQ3): Our results show that LLMs do differentiate idiomatic from literal usage. Monolingual probing and LogitLens reveal a clear separation from about layer 7, with English showing the strongest signals and Italian and Turkish weaker but consistent patterns. Cross-lingual probing confirms this, with English transferring well to both languages and Turkish performing poorly as

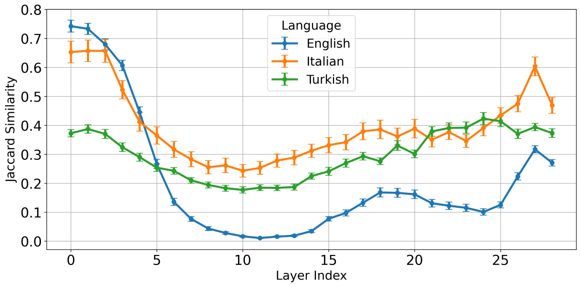

Figure 6: Mean Jaccard (LogitLens) by layer for PIEs on Llama-3.2-3B.

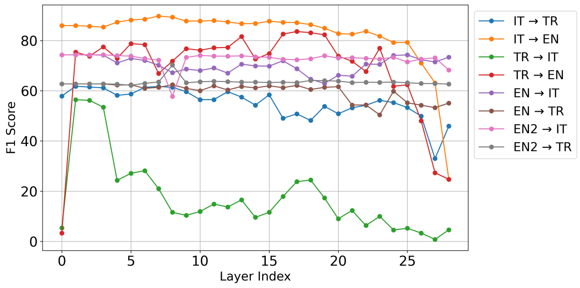

Figure 7: Cross-lingual linear-probe F1 across Llama-3.2-3B layers.

a source. Overall, idiomatic meaning is encoded in intermediate layers and is detectable within and across languages.

5 Conclusions

In this work, we studied cross-lingual idiom identification using both encoder- and decoder-based language models. We addressed three research questions examining (1) the ability of models to transfer idiom identification across languages, (2) the effect of cross-lingual transfer with language mixtures, and (3) how idiomatic meaning is represented internally.

Our experiments show that idiomatic meaning encodes transferable semantic patterns: encoder models achieve moderate cross-lingual generalization from a single source language, while decoder models show more selective transfer, especially from high-to-low-resource languages. Training with language mixtures further improves performance, with encoder models reaching near-baseline accuracy even when only a small portion of the mix comes from the target language, offering a practical approach for low-resource idiom identification. Probing analyses reveal a clear separation between idiomatic and literal usages in intermediate layers, with English exhibiting the strongest non-compositionality signals and Italian and Turkish showing weaker but consistent patterns. Together, these results demonstrate that LLMs capture idiomatic meaning in a way that supports cross-lingual transfer and can be leveraged to build scalable systems for multilingual idiom identification.

Limitations

Our study, although providing valuable insights into cross-lingual fine-tuning for idiomatic expression identification, is subject to several limitations. First, our reliance on existing datasets introduces variability in annotation definition and quality, as well as idiom coverage. Moreover, the number of samples and unique PIEs available, particularly for low-resource languages, varies and is limited in some cases, which may hinder robust learning. Second, the high computational cost of fine-tuning large generative models, such as Llama-3.1-70B, restricts the scope of hyperparameter tuning. For the same reason, the number of inference runs is limited. Both potentially contribute to the observed variability in performance and limit our ability to fully assess cross-lingual transfer.

Another limitation, noted in the explainability section, is that we were able to leverage only a subset of the data. While the findings present somewhat mixed signals, we view them as a promising indication that motivates further exploration into the factors enabling successful cross-lingual generalization of idiomatic meanings.

Ethics Statement

We utilized publicly available datasets and models in accordance with their licenses, which are detailed in Appendix I. No personally identifiable information was processed, and no new data was collected. Our work is intended solely for research purposes. We believe that our work poses no potential risks.

Acknowledgments

This study is supported in part by the European Research Council (Intellexus, Project No. 101118558). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them.

References

Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit, and Joakim Nivre. 2025. Using LLMs to advance idiom corpus construction. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

J Briskilal and C.N. Subalalitha. 2022. An ensemble model for classifying idioms and literal texts using bert and roberta. *Information Processing Management*, 59(1):102756.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, pages 19–22.

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.

GülŞen Eryiğit, Ali Şentaş, and Johanna Monti. 2022. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4):909–941.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 112–118, Berlin, Germany. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. Easy as PIE? identifying multi-word expressions with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23782–23801, Suzhou, China. Association for Computational Linguistics.

Chikara Hashimoto and Daisuke Kawahara. 2009. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation*, 43(4):355–384.

Wei He, Marco Idiart, Carolina Scarton, 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 Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent-BERT: Patent classification with fine-tuning a pretrained BERT model. *World Patent Information*, 61(101965).

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-GPT: Table fine-tuned GPT for diverse table tasks. *Proc. ACM Manag. Data*, 2(3).

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. Association for Computational Linguistics.

Marinela Parovic, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 176–193, Toronto, Canada. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. EPIE dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue*, pages 87–94, Cham. Springer International Publishing.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024. LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand. Association for Computational Linguistics.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar 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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Yuxin Zuo, Wenxuan Jiang, Wenxuan Liu, Zixuan Li, Long Bai, Hanbin Wang, Yutao Zeng, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2024. Alignxie: Improving multilingual information extraction by cross-lingual alignment. *arXiv e-prints*, pages arXiv–2411.

A Dodiom Dataset Construction

The original Dodiom dataset provides examples of idiomatic and literal usage for a predefined set of idioms in Turkish and Italian. Each example includes the sentence and the idiom it contains, but does not provide token-level span annotations.

To adapt the dataset for our idiom identification task, we convert the idiom-level annotations into token-level BIO tags. Since the surface form of each idiom is provided, we align it to the sentence using normalized character-level matching and tag its constituent tokens accordingly. Tokens belonging to the idiom are labeled as B (for the first token) or I (for subsequent tokens), while all other tokens are labeled as O.

As the original dataset does not include a predefined train-test split, we construct one to support reliable evaluation. Specifically, for each idiom, we randomly select two idiomatic and two literal examples (if available) for the test set. If fewer than two examples exist for a given class, we include a single instance. This strategy ensures that the test set includes both literal and idiomatic usages for as many idioms as possible, enabling balanced evaluation across expression types.

To reach a target size of 500 test examples, we first apply the above per-idiom sampling procedure. If the resulting set contains fewer than 500 examples, we then randomly sample additional examples, regardless of idiom identity or label, from the remaining pool to fill the gap. All remaining examples are assigned to the training set.

B MAGPIE Dataset Construction

MAGPIE was initially designed for the task of idiom *classification*, where each instance consists of a context made up of five sentences taken from a source document. The target phrase indicated by the idiom appears in the middle sentence, and the objective is to classify it as either idiomatic or literal. To adapt the dataset for the task of idiom *identification*, we merge all five sentences into one continuous span and label the target idiom using the BIO tagging scheme: B and I signify tokens that are part of the idiom. Conversely, O denotes all tokens that are not part of it.

This adaptation introduces potential noise, as additional idioms may appear in the surrounding context but remain unlabeled. In practice, we observe that models occasionally identify such expressions correctly, receiving false penalties due to missing gold annotations. Nevertheless, we consider this issue acceptable given its low frequency and limited impact on evaluation.

To minimize uncertainty, we eliminate 60 instances where the target PIE appears multiple times in the combined text, as it is not clear whether all occurrences should be labeled. The resulting adapted dataset comprises 4,391 test instances.

C Open MWE Dataset Construction

To adapt the corpus for idiom *identification*, we convert each example into the standard BIO tagging format. The dataset marks expressions using

LoRA	
r	64
α	16
Dropout	0.05
Training	
Epoch	3
Effective batch size	32 for Llama-8B, 8 for 70B
Learning rate	1e-5
Max grad norm	0.3

Table 2: LoRA and training hyperparameters used for fine-tuning.

angle brackets (“<” and “>”), and assigns each a label indicating figurative (`I`) or literal (`L`) usage. We retain only the idioms labeled `I`, and convert their constituents into BIO tags—labeling the first token as `B-IDIOM`, subsequent tokens as `I-IDIOM`, and all other tokens as `O`.

Additionally, the original dataset does not contain predefined training and test splits. To construct a reliable evaluation setting, we partition the data ourselves. For each idiom, we randomly select up to two idiomatic examples and two literal examples (if available) to form the test set. The remaining examples are used for training.

D Fine-Tuning

D.1 Decoder Fine-Tuning Hyperparameters

Training details are summarized in Table 2, with hyperparameters selected based on preliminary experiments.

D.2 Decoder Mix Fine-Tuning

The hyperparameters remained the same as those used in the standard fine-tuning setup (Table 2).

D.3 Encoder Fine-Tuning

The listed hyperparameters (Table 3) were selected based on preliminary experimentation; all remaining settings follow the default configuration of the `Trainer`[5] class in the `transformers` library.

For both mBERT and XLM-Roberta, we initially optimized hyperparameters on the English portion of ID10M and then reused the best configuration for the other languages. For mBERT, a learning rate of $2\text{e}{-}5$ with batch size 8 yielded the best results (F1 ≈ 0.76); with batch size 32 and 15 epochs we observed F1 ≈ 0.71; with batch size 128 and 25 epochs F1 ≈ 0.65; and with batch size 32 and 25 epochs F1 ≈ 0.76. For XLM-Roberta, we began with mBERT’s best hyperparameters, tested settings from “Using LLMs to Advance Idiom ...” which did not produce good results, and then conducted a grid search over learning rates $\{5\text{e}{-}6,\ 1\text{e}{-}5,\ 2\text{e}{-}5,\ 3\text{e}{-}5,\ 4\text{e}{-}5,\ 5\text{e}{-}5\}$ and batch sizes $\{8,\ 12,\ 16,\ 32\}$; the best result reached F1 ≈ 0.48.

Parameter	**XLM-RoBERTa**	**Other Models**
Epochs	20	20
Batch size	8	32
Learning rate	2e-5	2e-5

Table 3: Encoders fine-tuning hyper-parameters.

E Generative Decoder Evaluation

Following prior work on idiom identification, we evaluate model performance using the macro-averaged F1 score at the token level, comparing predicted BIO tags against gold annotations.

To enable this, we first convert the model’s predicted idioms into token-level BIO tags. For all languages except Japanese, we normalize the predicted idioms and input sentence (lowercasing, removing dashes and surrounding quotes, standardizing quotation characters), and apply character-level substring matching to align idioms with token spans. For JA, we use MeCab [6] to tokenize both the input and predicted idioms, ensuring alignment.

The evaluation pipeline includes a fallback mechanism for handling malformed or non-standard model outputs. When predictions cannot be parsed as valid JSON, we attempt recovery using relaxed parsing strategies such as key normalization (e.g., accepting “idioms”, “idoms”, “idiom”, etc.). If all recovery attempts fail, we treat the prediction as empty, labeling all tokens as ’O’.

All evaluation code, including normalization, parsing, fallback recovery, BIO tagging, and metric computation, is available in our project repository.

F Prompt Selection and Design

To identify an effective prompting strategy for fine-tuning, we experimented with multiple configurations before settling on the final schema used in our models. These initial explorations were motivated by both practical considerations (e.g., output

[5] https://huggingface.co/docs/transformers/en/main_classes/trainer

[6] https://pypi.org/project/mecab-python3/

consistency, ease of parsing) and alignment with prior work.

We tested two main design choices:

Instruction-based vs. conversational formatting. We compared flat instruction-style prompts using a `{"prompt": ..., "completion":..}` format (Figure 10) with conversational schemas that follow role-based formatting: `{"role":"system"|"user"|"assistant","content":..}`. We found that the conversational format—particularly when using a well-scoped system message—led to more structured and reliable outputs, especially for instruction-tuned Llama models.

Output format: free text vs. structured responses. We tested open-ended natural language completions versus *structured formats*, such as:

- A TVc-style list as proposed in CoAM (Ide et al., 2025) (e.g., `["spill the beans", "break the ice"]`)
- A JSON-based structure, such as:

```
{
  "idioms": ["spill the beans"],
}
```

See example for different prompt design in 10.

After thorough testing, we chose a conversational setup using JSON format; see the final prompt in Figure 9.

G Resources

Encoder-based models. We run our fine-tuning experiments on an NVIDIA GeForce RTX 3090 machine. Overall, all runs and tests took approximately 235 hours. The exact code and package versions required are published in the project's repository. [7]

Generative decoder models. We conduct both fine-tuning and inference using the Together AI platform. For fine-tuning, we do not explicitly select the hardware resource. For inference, however, we choose the `1 NVIDIA H100 80 GB SXM` hardware. The total cost of this setup is approximately $60.

Explainability analysis. To run this analysis, we use a single NVIDIA GeForce RTX 3090 machine with 24GB.

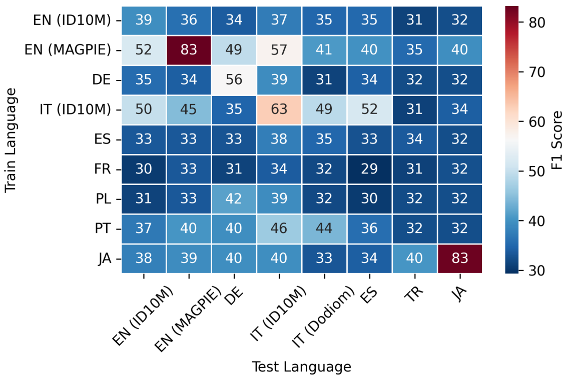

Figure 8: F1 score of the fine-tuned XLM-RoBERTa model. Each cell represents average macro F1 across five random seeds. For EN, both ID10M and MAGPIE are used for training and evaluation; for IT, we train on ID10M and evaluate on both ID10M and Dodiom.

H Full Results

H.1 Encoder heat map results

Figure 8 shows the heat map results for XLM-RoBERTa.

H.2 Encoder full results

Table 6 presents the complete set of evaluation results across all monolingual FT models, and the table has the results for the mix task 7

H.3 Generative Decoder

Table 8 presents the complete set of evaluation results across all models and configurations.

I License

We detail the models, datasets, and packages we use and their respective licenses in Table 4.

J Model Checkpoints

In Table 5, we present the checkpoints used in this work and the models' sizes.

K AI assistants

We utilized AI assistants, such as ChatGPT, to assist with code formatting, phrasing suggestions, and LaTeX styling during writing. All outputs were reviewed and edited by the authors, ensuring no content was generated or used without human verification.

[7] https://github.com/Intellexus-DSI/easy-as-pie

Model input

System prompt
You are a professional linguist specializing in figurative language, and your task is to analyse sentences that may contain an idiom, also known as an idiomatic expression. This is a definition of idiom: 'A phrase, expression, or group of words that has a meaning different from the individual meanings of the words themselves, and employed to convey ideas in a non-literal or metaphorical manner'. Mark idioms only when their usage in the context is idiomatic/figurative and let literal meanings remain unmarked.

User prompt
You are given one sentence in *language*, you are an expert of this language. If detected, write the idioms exactly as they are in the sentence, without any changes. Only answer in JSON.
Sentence: *Sentence*

Assistant
Output: *Expected out put*

Figure 9: Final Fine Tune message

Artifact	Type	License	Notes
Open MWE [a]	Dataset	BSD 3-Clause	subset was used
ID10M[b]	Dataset	CC BY-NC-SA 4.0	Full test set used
MAGPIE[c]	Dataset	CC BY 4.0	Used a filtered subset
Dodiom[d]	Dataset	GNU General Public License v3.0	subset was used
Llama-3.1-8B-Instruct-Reference[e]	Model	Llama 3.1 Community License	-
Llama-3.1-70B-Instruct-Reference[f]	Model	Llama 3.1 Community License	-
mBERT[g]	Model	apache-2.0	-
XLM-RoBERTa[h]	Model	MIT	-
LangChain[i]	Framework	MIT License	Used for prompting
Together AI[j]	Provider	Proprietary	Used for API access
NNsight[k]	package	MIT License	Used for logit lens analysis
scikit-learn [l]	package	BSD license	Used for Linear probing
MeCab[m]	package	BSD 3-Clause License	Used for JA tokenization

[a] https://github.com/nlp-waseda/OpenMWE
[b] https://github.com/Babelscape/ID10M/tree/master
[c] https://github.com/hslh/magpie-corpus/tree/master?tab=readme-ov-file
[d] https://github.com/Dodiom/Dodiom
[e] https://huggingface.co/meta-Llama/Llama-3.1-8B-Instruct
[f] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
[g] https://huggingface.co/google-bert/bert-base-multilingual-cased
[h] https://huggingface.co/FacebookAI/xlm-roberta-base
[i] https://www.langchain.com
[j] https://www.together.ai
[k] https://nnsight.net
[l] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[m] https://pypi.org/project/mecab-python3

Table 4: License and usage summary of all datasets, models, and tools used in this study.

Model	Checkpoint	# Parameters
Llama-3.1-8B	Meta-Llama-3.1-8B-Instruct-Reference	8B
Llama-3.1-70B	Meta-Llama-3.1-70B-Instruct-Reference	70B
XLM-RoBERTa	FacebookAI/xlm-roberta-base	279M
mBERT	google-bert/bert-base-multilingual-cased	179M

Table 5: Checkpoints used during experiments and the number of parameters.

Prompt Variant 1: Instruction-style Format (Not Used in Final Experiments)

Prompt:
You are a professional linguist specializing in figurative language and your task is to analyse sentences that may contain an idiom, also known as an idiomatic expression.
This is a definition of idiom: 'A phrase, expression, or group of words that has a meaning different from the individual meanings of the words themselves, and employed to convey ideas in a non-literal or metaphorical manner'. Identify all idioms in the following sentence, but only if they are used figuratively. Do not include literal usages. Return your answer in JSON format.
Sentence: `I spilled the beans during the meeting.`
Completion: `{"idioms": ["spilled the beans"]}`

Prompt Variant 2: Alternative Wording (Not Used in Final Experiments)

System prompt
You are a professional linguist specializing in figurative language and your task is to analyse sentences that may contain an idiom, also known as an idiomatic expression. This is a definition of idiom: 'A phrase, expression, or group of words that has a meaning different from the individual meanings of the words themselves, and employed to convey ideas in a non-literal or metaphorical manner'.

User prompt
You are given one sentence in a language, you are an expert in this language. Your task is to identify idioms only if they are used in an idiomatic or figurative sense. If the usage is literal, do not mark it.
Output only the idioms that appear exactly as they are in the sentence, without any changes.
Return the answer in JSON format only.
Sentence: `I spilled the beans in the kitchen.`

Assistant
`{"idioms": []}`

Figure 10: Prompt formatting strategies explored during fine-tuning

Test Lang. Train Lang.	Model	EN (ID10M)	EN (MAGPIE)	DE	IT (ID10M)	IT (Dodiom)	ES	TR	JA
EN (ID10M)	mBERT	75.30±0.50	54.30±1.53	56.27±0.90	61.75±1.43	53.22±0.75	59.08±1.01	33.20±1.23	35.19±0.73
	XLM-RoBERTa	39.49±14.54	35.87±6.06	33.72±4.66	37.46±10.68	34.79±7.14	35.09±10.24	31.32±0.39	32.36±0.00
EN (MAGPIE)	mBERT	63.51±3.66	80.94±2.50	57.52±1.80	61.86±0.95	53.22±1.71	52.85±2.89	41.71±4.16	42.86±1.24
	XLM-RoBERTa	52.02±5.73	83.25±0.97	49.34±2.38	57.32±4.93	40.72±5.43	40.02±5.18	35.40±1.30	39.61±2.40
DE	mBERT	43.11±0.32	37.24±2.11	76.22±2.33	50.70±1.09	35.26±1.18	41.63±1.23	38.53±0.97	36.19±2.00
	XLM-RoBERTa	35.41±2.91	34.11±0.59	56.27±6.20	39.24±3.46	31.30±0.82	34.24±2.79	32.10±1.17	32.35±0.00
IT (ID10M)	mBERT	47.93±0.51	42.23±0.77	39.96±3.48	68.79±0.73	53.12±0.53	55.23±1.89	36.29±1.27	36.80±0.93
	XLM-RoBERTa	49.81±1.48	44.61±1.18	34.71±0.81	62.98±2.63	48.97±2.33	51.83±1.46	31.17±0.54	34.04±1.10
ES	mBERT	36.76±0.47	38.09±0.82	44.61±2.23	54.78±1.96	37.30±0.84	37.63±0.43	34.22±1.06	32.48±0.16
	XLM-RoBERTa	33.43±2.21	33.39±0.56	32.94±3.07	37.55±3.53	34.79±2.11	32.77±2.24	33.52±1.72	32.36±0.00
JA	mBERT	32.45±0.52	34.11±0.51	35.41±0.87	35.90±1.02	32.46±0.50	33.09±1.19	42.95±2.36	82.92±1.21
	XLM-RoBERTa	37.89±2.03	39.21±1.53	39.61±3.29	39.56±3.29	33.27±1.96	34.00±3.69	40.29±3.53	82.68±2.02
FR	mBERT	30.64±0.54	33.78±0.33	32.72±0.90	35.57±1.24	32.66±1.48	31.42±0.30	31.29±0.22	32.71±0.28
	XLM-RoBERTa	30.39±0.01	33.39±0.32	31.25±0.10	33.80±1.64	31.99±1.20	29.29±0.03	31.14±0.00	32.50±0.14
PL	mBERT	32.50±1.11	33.66±0.23	44.04±4.28	41.25±0.34	30.51±0.27	32.22±0.72	33.62±1.26	32.40±0.04
	XLM-RoBERTa	31.31±0.35	33.30±0.27	41.67±0.36	39.39±0.78	31.99±0.20	30.37±0.58	32.19±0.33	32.36±0.00
PT	mBERT	34.06±1.21	41.98±0.32	44.58±1.86	51.79±1.75	44.64±1.08	38.03±0.73	34.21±1.44	34.82±0.37
	XLM-RoBERTa	37.05±4.89	40.33±4.56	39.60±5.67	46.38±9.28	44.03±7.55	35.75±4.49	32.29±0.75	32.40±0.10

Table 6: The performance of fine-tuned mBERT and XLM-RoBERTa was evaluated across eight languages. The results are reported in terms of F1 scores (bounds = mean ± standard deviation over five runs).

Language	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
EN	41.06±2.55	63.22±1.44	67.55±2.00	69.64±1.22	71.50±0.59	71.96±1.71	72.67±1.12	73.43±0.62	74.05±0.23	73.74±0.52	74.31±0.85
DE	65.18±2.32	70.80±0.67	71.34±1.69	72.32±1.76	72.87±2.44	73.24±2.27	73.05±1.83	75.62±2.96	76.47±2.09	74.64±0.89	75.56±2.04
IT	60.91±1.50	67.39±0.54	67.93±0.33	68.32±0.89	68.92±0.69	68.75±0.43	69.39±0.52	69.09±0.93	70.11±0.59	69.79±0.70	69.18±0.97
JA	33.16±0.30	47.70±1.08	49.38±0.74	50.09±1.02	50.32±0.87	51.04±1.36	51.80±1.04	52.85±0.69	52.44±0.99	52.48±1.21	52.45±0.94
ES	50.57±2.19	52.51±1.44	48.59±2.00	45.95±1.41	42.68±2.61	41.82±2.50	40.21±1.26	37.79±1.24	37.50±2.19	37.57±2.33	37.55±2.04

Table 7: The performance of fine-tuned mBERT mix.The results are reported in terms of F1 scores (bounds = mean ± standard deviation over five runs).

Test Lang.	Model	EN (ID10M)	EN-magpie	DE	IT (ID10M)	IT (Dodiom)	ES	TR	JA
Baseline	Llama 70B	81.31	68.21	77.68	73.11	58.49	69.07	51.87	46.9
	Llama 8B	81.94	53.52	72.14	64.51	51.09	65.09	46.32	42.38
EN (ID10M)	Llama 70B	81.34	67.07	67.06	63.62	58.26	61.82	47.53	50.85
	Llama 8B	80.14±1.37	53.31±0.82	71.80±0.09	64.26±1.59	51.52±0.34	65.53±0.45	46.41±0.42	42.79 ±0.15
DE	Llama 70B	81.6	67.03	75.88	71.74	57.53	70.34	50.67	51.14
	Llama 8B	82.17±0.75	53.22±0.28	71.50±0.31	64.46±0.22	50.66±0.75	65.35±1.32	46.46±0.40	42.82±0.44
IT (ID10M)	Llama 70B	81.45	67.17	76.41	71.84	57.44	70.19	50.82	50.59
	Llama 8B	81.14±0.89	52.98±0.50	70.13 ±0.83	63.75±0.75	51.17±0.25	65.42±0.43	46.55±0.27	42.57±0.29
ES	Llama 70B	81.34	59.75	77.04	71.07	57.68	70.26	50.65	50.45
	Llama 8B	81.37±0.56	53.06±0.08	71.02 ±0.43	64.14±0.79	51.39±0.51	64.65±0.73	46.34±0.94	42.08±0.15
TR	Llama 70B	81.64	68.45	77.93	72.62	58.13	69.93	51.41	50.87
	Llama 8B	81.18±0.61	53.15±0.55	71.37 ±0.55	65.09±0.27	50.80±0.46	65.54±0.36	46.31±0.93	42.32±0.17
NL	Llama 70B	81.47	67.24	76.95	71.11	58.02	70.22	50.1	50.78
	Llama 8B	80.43±1.02	52.78±0.24	71.60 ±0.24	64.47±0.55	51.35±0.66	64.98±0.75	47.41±0.87	42.80±0.75
FR	Llama 70B	81.23	67.49	76.83	70.59	58.19	70.08	49.74	50.64
	Llama 8B	81.49±0.52	53.09±0.16	72.48 ±0.94	64.72±0.97	50.74±0.26	64.55±1.26	46.58±0.46	42.53±0.17
PL	Llama 70B	81.21	67.37	77.6	71.43	58.01	70.38	49.88	50.39
	Llama 8B	81.43±0.39	53.09±0.57	71.32 ±1.01	64.13±1.81	50.93±0.33	65.89±0.51	46.20±1.04	42.03±0.51
PT	Llama 70B	81.34	66.39	76.79	71.45	57.22	69.64	50.27	50.88
	Llama 8B	81.84±0.97	52.70±0.44	71.45 ±1.50	64.29±0.25	50.87±0.29	65.42±0.8	46.57±0.25	42.50±0.56
Small mix	Llama 8B	81.77±0.88	53.10±0.43	63.70±2.66	62.48±1.46	51.78±0.32	65.72±0.73	46.59±0.73	43.43±1.93
Large mix	Llama 8B	79.07±0.62	55.44±0.66	46.98±0.68	51.75±1.05	52.45±0.56	65.86±1.16	42.94±1.93	41.17±0.56

Table 8: Performance of Fine-Tuned Generative Decoder Models on Test Languages. Reports F1 scores for Llama 8B and Llama 70B models fine-tuned on various language combinations. Results are shown across our test languages. Bounds denote mean ± standard deviation over five runs.