NLP Datasets for Idiom and Figurative Language Tasks

Blake Matheny[1], Phuong Minh Nguyen[1], Minh Le Nguyen[1], Stephanie Reynolds[2]

[1]Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan.
[2]General Education, International College of Technology, Kanazawa, Japan.

Contributing authors: matheny.blake@jaist.ac.jp; phuongnm@jaist.ac.jp; nguyenml@jaist.ac.jp; s-reynolds@neptune.kanazawa-it.ac.jp;

Abstract

Idiomatic and figurative language form a large portion of colloquial speech and writing. With social media, this informal language has become more easily observable to people and trainers of large language models (LLMs) alike. While the advantage of large corpora seems like the solution to all machine learning and Natural Language Processing (NLP) problems, idioms and figurative language continue to elude LLMs. Finetuning approaches are proving to be optimal, but better and larger datasets can help narrow this gap even further. The datasets presented in this paper provide one answer, while offering a diverse set of categories on which to build new models and develop new approaches. A selection of recent idiom and figurative language datasets were used to acquire a combined idiom list, which was used to retrieve context sequences from a large corpus. One large-scale dataset of potential idiomatic and figurative language expressions and two additional human-annotated datasets of definite idiomatic and figurative language expressions were created to evaluate the baseline ability of pre-trained language models in handling figurative meaning through idiom recognition (detection) tasks. The resulting datasets were post-processed for model agnostic training compatibility, utilized in training, and evaluated on slot labeling and sequence tagging.

Keywords: idiom, idiomaticity detection, data annotation, language models

1 Introduction

The datasets created in this paper are downstream derivations of the Common Crawl [1] corpus. Many datasets have been derived from Common Crawl and shared for research. Current datasets focusing on idioms or figurative language have many features, but offer limited context, limited size, or outdated vernacular. In an effort to create a new, relevant, and updated dataset of idiomatic expressions and figurative language representing more recent vernacular, in this research OSCAR [2] [1]and C4 [2] filters of Common Crawl were used to match a compiled list of idioms. As an homage to the original researchers of OSCAR and C4, ***p**otential* **i**diomatic and **f**igurative **l**anguage expressions (*P*IFL), and IFL-OSCAR-A and IFL-C4-A, two additional human-annotated datasets of definite idiomatic and figurative language expressions were developed and used as training data for the task of idiomaticity detection.

These idiomatic and figurative expressions are ubiquitous in natural language, but defy straightforward compositional interpretation. For language models, expressions like *neither fish nor fowl* cannot be accurately understood by interpreting their individual words in isolation. As such, they pose significant challenges for NLP systems in tasks ranging from machine translation to semantic similarity and information retrieval.

Several idiomatic expression datasets have been introduced in recent years. LIdioms [3], EPIE[4], [5], FLUTE [6], and SemEval-2022 [7] provide annotated examples for various figurative language types, yet they differ in size, labeling schema, context availability, and idiom diversity. Consequently, a dataset was derived from OSCAR's filter of a snapshot of 2018 Common Crawl using phrase vocabulary combined from these datasets. The *P*IFL-OSCAR dataset contains 5.8 million examples and 3,207 unique idioms. Social media has contributed to significant changes in the English lexicon. The increased in informality has begun permeating formal discourse. As idioms and figurative language appear more frequently during informal exchanges rather than formal situations, they appear more frequently in messages and thereby in lexicon used with AI chats. It is necessary to include as much current language as possible when evaluating and training for these sociolinguistic phenomena. In this work, our contributions are:

1. One large pre-sampled dataset for generalization.
2. Two new annotated datasets with a larger variety of labels and metadata[3].
3. Cross evaluation on existing datasets.
4. SOTA performance of sequence accuracy metric on MAGPIE.

Natural language understanding involves not only predicting words (token prediction) but interpreting roles (token class prediction), particularly when encountering figurative expressions. Idioms, which frequently defy compositional logic, require models to recognize spans that signal non-literal meaning and then assign them appropriate semantic labels or types. These idiomatic expressions are challenging due to their often

[1] https://creativecommons.org/publicdomain/zero/1.0/
[2] https://oscar-project.org/
[3] The datasets and source code will be published upon acceptance.

unpredictable alignment with surface syntax and their cultural or contextual dependencies. Due to the abstract nature of many multi-word expressions which are typically non-compositional in nature, the combined semanticity is not equal to the parts/-words in the phrase. If the single meaning of a phrase is known, any small change in idiom phrasing will cause a language model to misclassify, misgenerate, or mistranslate (i.e., "trip up"). A native speaker will, according to Sinclair's Idiom Principle [8], subconsciously excuse the incorrect wording and recognize the speaker's or writer's intended meaning. These small changes can be single word mistakes ("two birds with one rock/stone"), or they can take the form of figures of speech errors, such as a mixed metaphor or malapropism, which can even be difficult for native speakers. Common malapropisms include "one in the same" in the place of "one and the same", "deep-seeded" for "deep seated", and "could care less" for "couldn't care less." A mixed metaphor combines parts of two or more common expressions in an unusual way. An example of a mixed metaphor that has been observed in films is as follows, "people in glass houses sink ships." The first part of this metaphor explains, "people in glass houses shouldn't throw stones," encouraging one not to criticize others for their faults. While the second part, "loose lips sink ships," means to beware of unguarded talk. Humans tolerate and understand these mistakes, while a language model may not be robust to them.

2 Related Work

This section represents a general overview of recent high-performing selection of models and methods, not a complete historical and comprehensive review of research on idiom and figurative language tasks and NLP applications. Early improvements for phrase-level representations in BERT [9] focused on contrastive learning and masked-token tasks. Wang et al. introduced Phrase-BERT, which fine-tunes BERT with a paraphrase-based contrastive objective to yield embeddings that better capture phrasal semantics and support applications like neural topic models through nearest-neighbor phrase retrieval [10]. Qin et al. proposed IBERT , framing idiom detection as a cloze [11]. Similarly, Zhou et al. tackled idiomatic expression paraphrasing without strong supervision by combining unsupervised contextual signals with back-translation to build large-scale parallel idiom–literal corpora, showing significant gains in BLEU, METEOR, and SARI for idiomatic sentence rewriting [12].

A parallel approach leveraged contextualized word embeddings for disambiguation. Škvorc et al. introduced MICE, using ELMo [13] and BERT embeddings as inputs to a neural classifier trained on a novel idiom dataset; they demonstrated cross-lingual transfer and strong PIE detection even for unseen expressions [14]. Kurfalı and Östling treated each candidate MWE as a single token and applied supervised and unsupervised classifiers over CWEs, achieving notable literal vs. idiomatic classification gains without extensive feature engineering [15]. A STARSEM'22 study further showed that integrating static, masked-token, and contextual embeddings from MLMs significantly boosts token-level idiom classification in multilingual settings [16]. Building on these insights, Yayavaram et al. introduced multi-objective fine-tuning of BERT with word

cohesion and machine-translation-based contrastive losses, yielding state-of-the-art sequence detection across seven diverse idiomaticity datasets [17].

More recently, large language models (LLMs) have pushed re-evaluation of idiomaticity detection. Phelps et al. benchmarked GPT-3, PaLM, and other LLMs on SemEval 2022, FLUTE, and MAGPIE, finding that while LLMs offer few-shot competitiveness, they consistently trail specialized, fine-tuned encoders [18]. De Luca Fornaciari et al. introduced IdioTS, a curated conversational LLM test suite for rare and novel idioms, revealing imperfect handling of unseen MWEs despite few-shot prompting strengths [19]. Wu et al. analyzed idiomatic clusters in dialogue-optimized LLM embeddings, showing more diffuse idiom representations compared to literal paraphrases and highlighting the need for embedding specialization [20].

Specialized objectives have continued to advance well for figurative expressions: He et al. developed an adaptive contrastive triplet loss that mines hard positive/negative idiom samples, boosting SemEval 2022 Task 2 performance with asymmetric distance learning [21]. Zeng and Bhat modeled idiomaticity as semantic compatibility between expression and context, using an attention-flow mechanism to fuse lexical and embedding features for robust PIE identification [22]. Madabushi et al. also released AStitchInLanguageModels, a bilingual MWE dataset with fine-grained senses, demonstrating limited zero-shot but sample-efficient few-shot idiomaticity detection and representation learning [23].

3 Current Datasets

High-quality datasets are contributing to these advances: FLUTE frames idiomaticity as an NLI task with 1.7K idiom pairs explained via textual entailment [6]; MAGPIE provides 56K crowd-annotated PIE instances from the BNC [5]; the EPIE static corpus supports cloze tasks and reading-comprehension models [11]; LIDIOMS links multilingual idioms in RDF [3]; and SemEval 2022 Task 2 delivers aligned detection and embedding splits in multiple languages [7]. VNC [24] was previously a common benchmark for idioms, though after examination and consideration, it was omitted from this research. Surveying model behavior, Miletić and Schulte im Walde show that transformer encoders inconsistently capture MWE (multi-word expression) semantics—often relying on surface patterns and localizing semantics in early layers, revealing a lack of task specific pretraining ability[25]. He et al. propose model-agnostic probing metrics (Affinity and Scaled Similarity) on noun-compound minimal pairs, revealing persistent gaps in idiomaticity representation across static and contextual models [21]. *P*IFL-OSCAR and the subsequent samples have also been developed to offer new benchmarks for model development and evaluation.

The datasets utilized in this research represent outstanding work with unique annotation and compilation methods. In an effort to enhance idiom and figurative language research with the most updated type of language, Common Crawl filters were used. This not only diversifies the discourse type, but includes current vernacular that will more likely contain figurative language. EPIE and FLUTE show the most recent data. However, closer examination of the EPIE corpus revealed a large quantity of archaic

Table 1 Summary of major idiom and figurative language datasets.

Dataset	Discourse / Source Texts	Date Range	Annotation Type	Task / Training Setup	# Idioms / Types	# Entries / Instances
MAGPIE	BNC: spoken, fiction, news, academic, non-fiction	1960s–1993	Crowdsourced annotations for idiomatic vs literal vs other usage	Binary / multi-class idiomaticity detection; random vs type-based splits	∼1,756	∼56,622
VNC-Tokens	BNC (verb–noun combos in sentences)	1960s–1993	Expert linguists, then validated	Binary classification (literal vs idiomatic)	53	2,984
SemEval-2022 Task 2	English (news, literature), Portuguese (OPUS), Galician (lit/news)	1990s–2010s	Crowdsourced, multilingual	Subtask A: idiomaticity classification; Subtask B: embeddings	English: 60; Portuguese: 50; Galician: 50	∼6,000 (all languages)
EPIE	BNC, COCA, newswire, conversational	1960s–2019	Human annotation (literal vs idiomatic)	Sequence labeling: idiomatic vs literal spans	717	∼18,000–20,000
FLUTE	Curated figurative texts: news, stories, online	2010s–2020s	Mix of crowdworkers + experts; explanation annotations	NLI framing (sentence pair entailment with figurative explanation)	4 categories (idioms, metaphors, similes, sarcasm)	∼9,000

syntax and lexicon. Since the direction of this research was recent contemporary lexicon, the EPIE corpus and dataset were omitted from training and testing, though its idiom vocabulary remained a principle component in building the new datasets.

The existing state-of-the-art results for each relevant dataset are seen in Table 2. With the exception of the sentence-pair entailment task in FLUTE, MAGPIE is the most recently benchmarked dataset. MAGPIE is split into magpie_random and magpie_type_aware. The magpie_random dataset was used in this research as a benchmark. FLUTE and SemEval 2022 Task 2 have been modified to be trained and evaluated with the slot tagging architecture in this paper. Only BIO labels were added in a new category, given the target phrase existence in the entry sequence. In Table 3, the diverse categories of each dataset can be seen. With the exception of the specific figurative language type that exists only in FLUTE, the *P*IFL-OSCAR and *P*IFL-C4 datasets contain all diverse categories, intended for downstream model agnostic training and benchmarking.

Table 2 Current SOTA of idiom and figurative language datasets. The asterisk (*) denotes that the result is taken from the corresponding dataset paper.

Dataset	**Task**	**Primary**	**Best**
EPIE (Formal)	Span tagging (BIO)	Acc.	**98.00** *
MAGPIE (random)	Span id. (exact)	Sequence Accuracy	**91.51** [17]
MAGPIE (type-aware)	Span id. (unseen types)	Sequence Accuracy	**70.47** [26]
SemEval-2022 2A	Idiomaticity existence	Macro-F1 (All)	**93.85** *
FLUTE (Idioms)	Sentence-pair entailment	Accuracy	**79.20** *

4 Annotating Methodology

The datasets introduced in this section contribute to the other primary resources for evaluating idiomatic and figurative language phenomena. Each dataset constructed was derived through a uniform semi-automated pipeline involving idiom retrieval, sentence extraction, and label augmentation with syntactic and semantic information. Post-processing steps incorporated POS tagging, BIO sequence labeling, cosine similarity metrics from BERT embeddings via the Span Search Algorithm, and contextual relationships across adjacent sentences. These enriched linguistic features enable diverse analytical and modeling approaches, including sequence tagging, contextual disambiguation, and representation learning. Following dataset construction, human annotation refined subsets of the data to ensure high-quality gold labels for figurative versus literal interpretation. The following subsections detail the composition and refinement of the *P*IFL-OSCAR and *P*IFL-C4 datasets into IFL-OSCAR-A and IFL-C4-A, while the results emphasize their scalability, linguistic diversity, and suitability for advancing research in idiom recognition and figurative language modeling.

Dataset	**prev/ next sentence**	**paraphrase**	**BIO tags**	**POS tags**	**phrase type guess**	**length**	**position in sent**	**similarity metrics**	**metadata**	**fig/lit label**	**Idiom Type Label**	**Human Annotated**
FLUTE	no	yes	no	no	no	yes	no	no	no	yes (all fig)	yes	yes
FLUTE + BIO (our addition)	no	yes	yes	no	no	yes	no	no	no	yes (all fig)	(as matched)	yes
SemEval 2022	yes	no	no	no	no	no	no	no	no	yes (existence)	no	yes
SemEval 2022 + BIO (our addition)	yes	no	yes	no	yes	yes	yes	no	no	yes (existence)	(as matched)	yes
MAGPIE	yes	no	yes (0,1 only)	yes	yes	yes	yes	no	no	yes	no	(crowdsourced)
OSCAR*P*IFL (ours)	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	no	no (direct match, possibles)
OSCAR_IFL_A (ours)	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	no	yes
C4_IFL_A (ours)	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	no	yes

4.1 Pipeline Framework

The raw dataset construction consisted of idiom and figurative language phrase vocabulary retrieval, Common Crawl filter choice, language choice, phrase matching, sentence retrieval, and data compilation. Subsequent post-processing steps were performed to augment the raw data with POS tags, BIO labels, various statistics, and pre-computed similarities using BERT output embeddings. This completed the *P*IFL-OSCAR dataset. After this semi-automated construction, human annotation was performed on a small sample. Beginning with a portion of the idiom list, it proceeded as follows and can be seen in Fig. 1:

- Step 1: Collect idiom and figurative phrase vocabulary
- Step 2: Identify viable corpus
- Step 3: Retrieve raw target sentences and context via matching
- Step 4: Post process: cosine similarities via BERT output embeddings, POS tags, initial BIO labels
- Step 5: Classify a phrase as "Sometimes" or "Always" an idiom or figurative language
- Step 6: Choose "Literal" or "Figurative" for a sample of entries in the "Sometimes" set

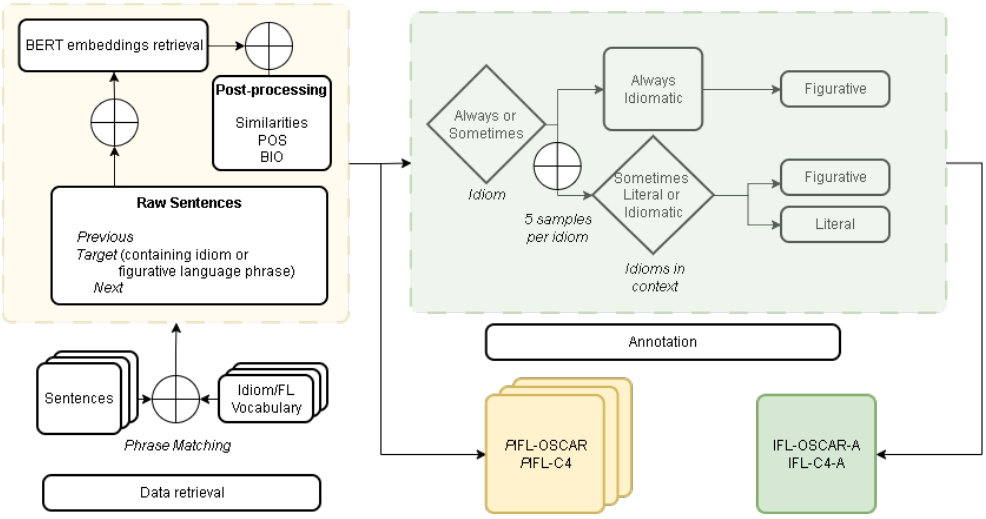

Fig. 1 Data retrieval and annotation structure.

4.2 Annotated Datasets

PIFL-OSCAR

The *P*IFL-OSCAR dataset was built from a list of idioms derived from MAGPIE, FLUTE, EPIE, and LIdioms and matched with the OSCAR filter of the 2018 snapshot of Common Crawl. All were presumed to be possible idiomatic or figurative language. The number of idioms amounted to 5,789, though 3,209 matches occurred during initial extraction for the dataset. It initially consisted of 98 files of 7.7 million target sentences, amounting to 21 GB. The data and metrics include POS tags for linguistic prior integration, BIO labels for sequence labeling, previous and next sentences for unsupervised learning, and many similarity metrics acquired from BERT output embeddings for post processing or supervised approaches. The dataset was further sanitized through keyword search to reduce specifically sexual, political, or discriminatory language, leaving a functional total of 5.83 million entries. Additionally, the list of unique idioms was reduced by the combination of similar phrases: "to be chuffed" = "be chuffed". By providing 97 subsets of the full dataset, it maintains the flexibility of training with samples. A sample of four subsets, named by the number of unique idioms and file number, were compared by idiom count distribution, shown in Fig. 2. This ensures that a single subset is an accurate sample of the full dataset of 97.

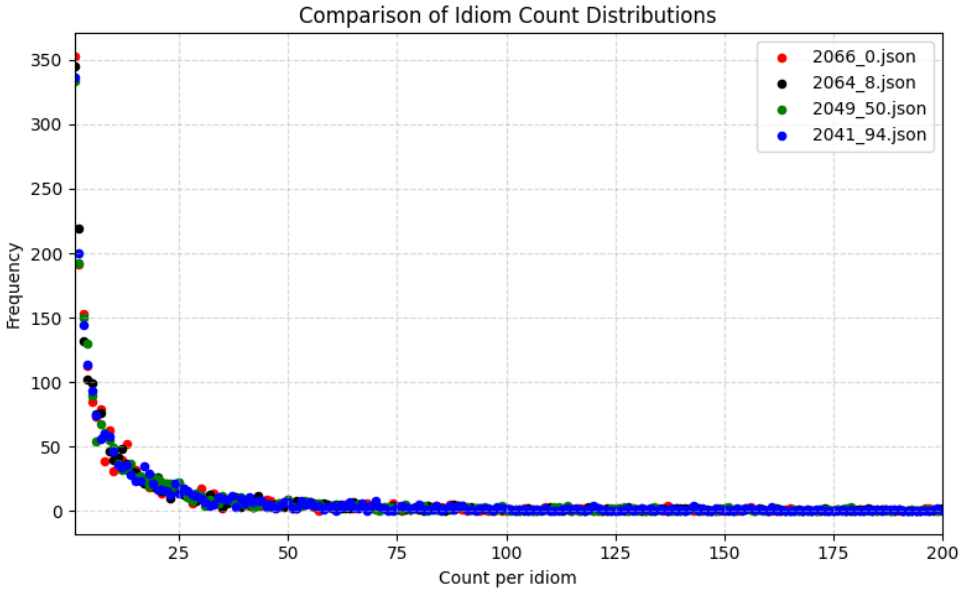

Fig. 2 File distribution comparison showing a single file is a suitable sample of the collection.

PIFL-C4

A similar end-to-end extraction was performed, as illustrated in Fig. 1 for the C4-filter [2] of the Common Crawl dataset[4]. This included phrase matching, BERT embedding processing, sanitizing via keyword, and combination of similar idioms. The subsequent annotation was also performed to acquire IFL-C4-A, though *P*IFL-C4 was not used in testing due to its smaller size and its similar blanket figurativeness to *P*IFL-OSCAR.

4.3 Span Search Algorithm

A re-imagining of an n-gram search algorithm was developed during preliminary testing of cosine similarity and figurative language testing. It created a set of phrases from a sentence. Such phrases were permutations of all sizes beginning at all positions. They were then ranked by the lowest cosine similarity of their output embeddings. For any Seq2Seq [27] or masked language model that is designed to output token embeddings, the algorithm is:

For each sentence, s, of length T with words, w_t, precompute its mean embedding

$$E_s = \frac{1}{T} \sum_{t=1}^{T} E_{w_t}.$$

We then slide a window of every length $\ell \in [2, L_{\max}]$ across positions $p = 1, \ldots, T - \ell + 1$. Each span (p, ℓ) has a phrase embedding

$$E_p = \frac{1}{\ell} \sum_{i=0}^{\ell-1} E_{w_{p+i}},$$

and a remainder embedding by subtracting the span's contribution component-wise,

$$E_r = \frac{T\,E_s - \ell\,E_p}{T - \ell}.$$

Compute two cosine similarities,

$$C_F = \cos(E_s, E_p), \quad C_R = \cos(E_r, E_p)$$

Preliminary testing showed a rank accuracy of exact phrase match for BERT-based models up to 10 percent as ranked by the lowest cosine similarity ranking of C_R (phrase/sentence remainder). Augmenting the datasets with these types of pre-computed values allows future researchers to determine threshold values or design models based on these values. Previous and next sentence inclusion with their respective similarities to the target sentence align with SemEval idiom existence tasks and BERT next sentence prediction objectives. Conceptually, the cosine similarities that have the most influence on future target-sequence models are the phrase/full sentence and the phrase/sentence remainder values. These elucidate the contextual effect of the

[4] https://huggingface.co/datasets/allenai/c4

phrase on the sentence. In combination with the labels existing in current datasets, the post-processed data presents the ability to train a variety of models using:

1. **BIO** - for sequence tagging, NER
2. **POS** - for parsing trees, linguistic priors
3. **Prev/Next**- for NSP (Next Sentence Prediction)
4. **Similarity** - for linguistic priors, composite scoring, candidate ranking
5. **Statistics** - n-gram modeling [5]

IFL-OSCAR-A and IFL-C4-A - Human Evaluation Most datasets established as benchmarks are heavily human annotated. For these datasets, two native-English-speaking linguists annotated figurative and literal labels for a subset of a single file of the *P*IFL-OSCAR dataset. The IFL-C4-A dataset sample was annotated by this author only after confirming alignment with the other annotator for IFL-OSCAR-A. The human evaluation was implemented via a voting application developed in Gradio[6]. Following the annotation pipeline, the initial votes on 300 idioms for the OSCAR-derived portion were voted "Always" and "Sometimes". Given two annotators choosing binary labels, Cohen's kappa coefficient was chosen to determine their alignment. Initial alignment was 82.6 percent. "Always" votes were removed from the list. Next, 5 examples of each remaining "Sometimes" idiom were extracted from 3 separate files to ensure randomness. These were voted "figurative", "literal", "unsure", or "omit". The omit choice was included to account for specific entries that were not able to be filtered by the keyword sanitizing effort. Business names, offensive or discriminatory language, urls, and first-name personal references are examples of sentences that might be omitted. The "unsure" category and any disagreements were reviewed by both annotators and resolved. The dataset is a result of this process. After reviewing entries, the agreement was 100 percent. Table 3 shows the final statistics of the datasets after omissions. IFL-OSCAR-A amounted to 695 entries with 152 unique idioms, while the IFL-C4-A contains 600 entries and 51 unique idioms.

Table 3 IFL-OSCAR-A statistics.

Totals	File 1 figurative 493	File 1 literal 210	File 2 figurative 485	File 2 literal 249
Overlapping entries	700	agree: 646	disagree: 54	Cohen's kappa: 82.36
After alignment (and 3 omissions)		Figurative: 472	Literal: 225	

5 Experiments

5.1 Datasets

The datasets created in this pipeline contain a total of 1300 annotated examples and a substantial 5.8 million *possible* entries, seen in Table 4. Not all examples in the larger

[5] A sample entry in Appendix 1.
[6] https://www.gradio.app/

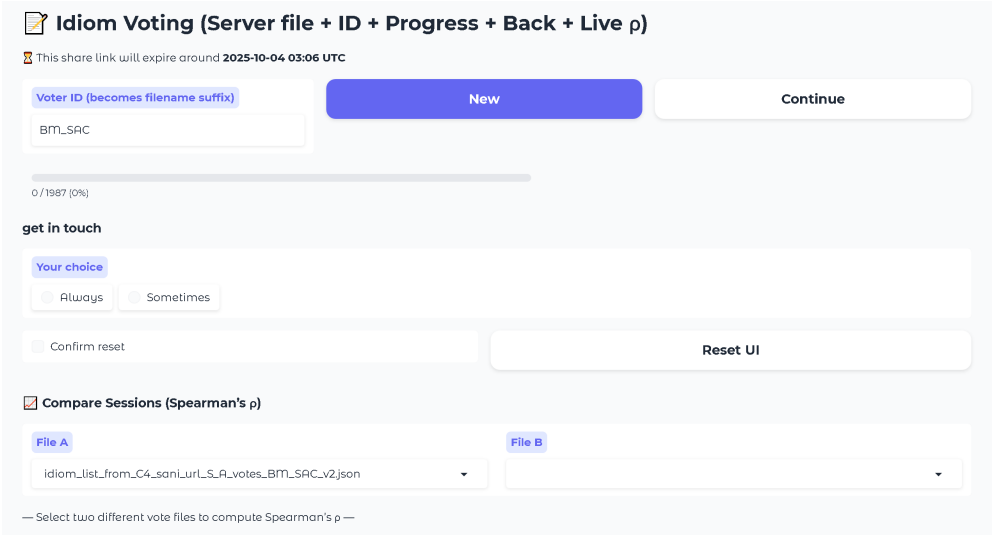

Fig. 3 "Sometimes" and "always" initial annotation.

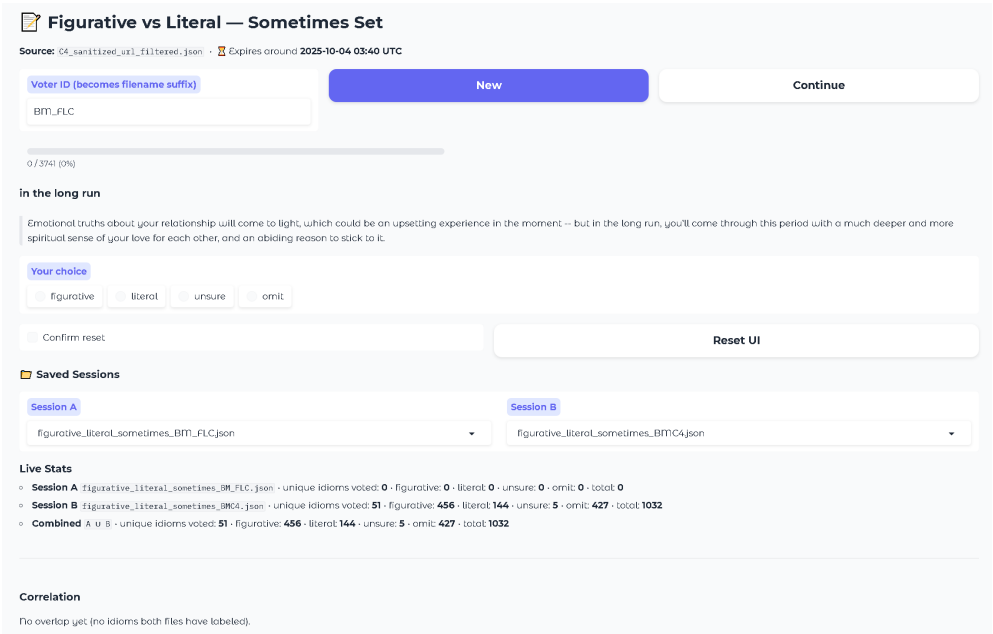

Fig. 4 "Figurative" and "literal" annotation example.

datasets are figurative, though later discussion will show that it retains important utility.

Table 4 Newly developed datasets and general statistics.

Dataset	Entries	Unique Idioms	Figurative	Literal
*P*IFL-OSCAR	5,834,942	3152	*(possible IFL)* 5,834,942	0
*P*IFL-OSCAR-HE	700	152	471	226
C4-PIFL	3,741	728	*(possible IFL)* 3,741	0
C4-PIFL-HE	600	51	456	144

The format of the newly created datasets contains a larger range of pertinent labels. The variety is intended for model-agnostic training and includes precomputed semantic relationships in addition to syntactical and statistical data. An example entry can be seen in Table 5

5.2 Experimental Models

Baseline testing of these datasets was conducted with a sequence labeling architecture. As a labeling architecture, it relied on parallel files of words and labels. An additional file containing a set of labels was required for word label prediction. The architecture, whose code base was re-implemented from JointBERT CAE [28] via Chen, et al [29] for joint loss modeling, was set up to accept various types of BERT-based models and a selection of other LLMs.

BERT-based Method.

The BERT-based model architecture comprises multiple Transformer encoder layers. Each encoder layer contains a self-attention mechanism that serves as the primary component for extracting and modeling linguistic features. This mechanism enables the model to capture long-range dependencies between word pairs within a sentence. Given an input sentence ($\mathbf{s} = \{w_i\}_1^{|s|}$ where $|s|$ denotes the number of words[7], the BERT model encodes each token into its corresponding hidden vector representation ($\mathbf{h}_i$).

$$\mathbf{h}^{[CLS]}, \mathbf{h}_i^{word} = \text{BERT}(\mathbf{s}) \tag{1}$$

Following Devlin et al. [9], the idiom recognition task is approached using a sequence labeling architecture (see Fig. 5). After obtaining the contextualized word representations $\mathbf{h}_i^{word}$ from the BERT encoder, each vector is passed through a fully connected (Dense) layer and a *softmax* classifier:

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}\mathbf{h}_i^{word} + \mathbf{b}) \tag{2}$$

[7] in practice, the input is tokenized into subwords using the BERT tokenizer.

Table 5 One data entry from the *P*IFL-OSCAR dataset.


```
{
    "labeling": {
        "seq.in": "Investing a little bit of time into research can save you a lot of money and trouble in the long run , and that is why you need to read this section carefully .",
        "seq.out": "O O O O O O O O O O O O O O O O O B-idiom I-idiom I-idiom I-idiom O O O O O O O O O O O O O O",
        "BIO": ["O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B", "I", "I", "I", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"],
        "seq.pos": "VBG DT JJ NN IN NN IN NN MD VB PRP DT NN IN NN CC NN IN DT JJ NN , CC DT VBZ WRB PRP VBP TO VB DT NN RB .",
        "idiom_POS": "IN"
    },
    "sentence_similarities": {
        "target_prev": 0.724046027407563,
        "target_next": 0.7110994006591008
    },
    "similarity_metrics": {
        "pos_pattern": "IN DT JJ NN",
        "phrase_type_guess": "prepositional phrase",
        "phrase_tokens": ["in", "the", "long", "run"],
        "sentence_tokens": ["investing", "a", "little", "bit", "of", "time", "into", "research", "can", "save", "you", "a", "lot", "of", "money", "and", "trouble", "in", "the", "long", "run", "and", "that", "is", "why", "you", "need", "to", "read", "this", "section", "carefully"],
        "idiom_token_span": [ 17, 18, 19, 20],
        "cosine_phrase_sentence": 0.7851315550686879,
        "cosine_phrase_minus_phrase": 0.7110709750649673,
        "cosine_phrase_left": 0.7259998505474241,
        "cosine_phrase_right": 0.5783826655791272,
        "cosine_phrase_local": 0.74453730902595,
        "cosine_delta_main": 0.07406058000372062,
        "contrast_delta_left": 0.05913170452126382,
        "contrast_delta_right": 0.20674888948956072,
        "idiomaticity_score_composite": 0.10350043850456644,
        "phrase_sentence_cosine": 0.5259013175964355,
        "backend": "hf-bert",
        "model_name": "bert-base-uncased"
    },
    "statistics": {
        "sentence_token_count": 32,
        "idiom_token_count": 4,
        "idiom_position_start": 17,
        "idiom_position_ratio": 0.53125,
        "ratio_idiom_to_sentence_length": 0.125,
        "percentage_idiom_position_in_sentence": 53.125
    },
    "line_index": 15,
    "model": {
        "name": "bert-base-uncased",
        "embedding_layer_type": "last_4_mean"
    },
    "match_type": "exact string match",
    "phrase": "in the long run",
    "phrase_canonical": "in the long run",
    "phrase_variants": [
        "in the long run"
    ],
    "phrase_representative": "in the long run"
}
```

where $\mathbf{W}, \mathbf{b}$ is learnable parameters. When a word is segmented into multiple subwords by the BERT tokenizer, only the first subword representation is used to represent the entire original word. For model training, the objective function is defined as the weighted sum of cross-entropy losses across all tokens in the sequence: losses of SF and ID sub-tasks.

$$\mathcal{L} = \frac{1}{|s|} \sum_{i=1}^{|s|} \text{CrossEntropy}(\hat{\mathbf{y}}_i, \mathbf{y}_i) \tag{3}$$

where $\mathbf{y}_i$ is the gold labels from Idiom Recognition datasets.

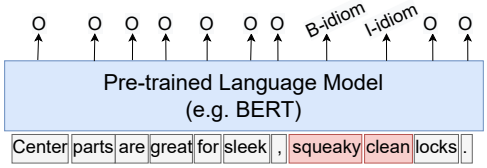

Fig. 5 BERT-based model utilizing sequence classification for idiom recognition tasks. The red highlight indicates the idiom phrase in the sentence.

LLM-based Method.

The ability of LLMs in the idiom recognition task was evaluated using a zero-shot instruction prompting technique. The selected LLMs were based on a Transformer decoder–only architecture. These models include Gemma-3, Llama-3.1, GPT-OSS, Mistral, Phi-4, and Qwen-2.5. Baseline testing was also performed on large chat-based LLMs via Ollama [8]. Given a sentence, the prompting template shown in Table 6 was used to instruct the LLM to understand and identify the idiom phrase in the sentence.

$$\text{output} = \text{LLM}(\text{prompting}(\mathbf{s})) \tag{4}$$

where *LLM* denotes the corresponding language model, and the *prompting* procedure applies the instruction template to explain the idiom recognition task to the model for a given input sentence $\mathbf{s}$.

5.3 Experimental Settings

The subsequent baseline experiments were tested using a 40GB A100 GPU. Batch sizes were tested and set at 4 for each run for maximum accuracy and consistency. Evaluation and logging steps were also set to reflect optimization: 10 for the much smaller *definite* IFL human evaluated model(s) and 1000 for the large *potential* IFL dataset. All models were cross evaluated on all datasets to test generalizability.

[8] Open-weight models (Gemma-3, Llama-3.1, GPT-OSS, Mistral, Phi-4, and Qwen-2.5) were accessed locally through the *Ollama* framework (https://ollama.com). See model-specific citations [30–34].

Table 6 Instruction prompting template for LLM-based idiom recognition tasks.

You are a slot tagger. Use ONLY the following tag types: Use ONLY prefixes 'B' and 'I'. If a token is outside, use 'O'. Return exactly N tags (one per input token), space-separated, no punctuation, no explanations. {{`sentence content, s`}}

5.4 Evaluation Metrics

The metrics, such as those shown in Table 2, are commonly used for evaluation of language models. Precision and recall help researchers use the true positives, false positives and false negatives to measure the effectiveness of models on given data and task. The F1 score formulates the combination of the precision and recall measurements, thereby giving an overall score for a given model. Sequence accuracy remains an equally important metric for inclusion, since idioms and figurative language are more often multi-word expressions.

Sequence Accuracy

Sequence accuracy is calculated by for slot labeling as the average of sequences, s_i, with all labels correctly assigned. For N number of labels The simple formula is calculated as:

$$\text{Sequence\_Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{\mathbf{s}}_i = \mathbf{s}_i\} \tag{5}$$

Entity-level performance

Precision, recall and F1 scores are basic scoring metrics for language models.

$$\text{TP} = \sum_{i=1}^{N} \big|\, E_i \cap \hat{E}_i \,\big|, \tag{6}$$

$$\text{FP} = \sum_{i=1}^{N} \big|\, \hat{E}_i \setminus E_i \,\big|, \tag{7}$$

$$\text{FN} = \sum_{i=1}^{N} \big|\, E_i \setminus \hat{E}_i \,\big|. \tag{8}$$

The precision measurement focuses on the relationship between all positive results predictions. All relevant instances are divided by all retrieved instances. For **TP**, true

positives and **FP**, false positives:

$$\text{Precision } (P) = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

Recall emphasizes only relevant instances by replacing false positives with **FN**, false negatives. It is measured by the following:

$$\text{Recall } (R) = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

While recall and precision give empirical impressions of model performance on specific data, the F1 score shows a model's general ability on the data. It is formulated with precision, **P**, and recall, **R**, as follows:

$$\text{F1} = \frac{2PR}{P + R}. \tag{11}$$

5.5 Results and Analysis

To establish additional necessity of this research and accurate baselines, the following open-weight chat LLMs were tested: Gemma-3, GPT-OSS, Llama3.1, Mistral, Phi-4, and Qwen2.5. The models were accessed and prompted via Ollama. Regardless of model size, the standard quantization on this platform is 4-bit, which helps the models perform more quickly, but decreases accuracy. Models were evaluated on slot precision, recall, F1, and sequence accuracy. The results for each test shown in Tables 7,8, 9 were generally poor except the sequence accuracy results in 10 for the SemEval 2022 dataset. All results for chat LLMs, automatically quantized, were averaged by metric and separated by dataset in 6.

Table 7 Results for test_slot_precision

Dataset	Gemma3	GPT-OSS	Llama3.1	Mistral	Phi4	Qwen2.5
IFL-C4-A	0.01	0	0	0.01	0.01	0
FLUTE	0.02	0	0	0.01	0.04	0.02
PIFL-OSCARPIE	0.01	1	0	0.01	0.01	0.01
IFL-OSCARPIE-A	0	0	0	0	0	0.03
SemEval-2022	0	0	0	0	0.01	0
magpie_random	0	1	0	0	0.02	0.01

5.5.1 LLM errors

Error analysis of the chat LLMs on a one-shot prompt setting show highly inaccurate predictions. A common error for each model within each dataset was chosen for analysis. Tables 11, 12 show a variety of incorrect labeling schemes. The types

Table 8 Results for test_recall

Dataset	Gemma3	GPT-OSS	Llama3.1	Mistral	Phi4	Qwen2.5
IFL-C4-A	0.04	0	0	0.04	0.02	0
FLUTE	0.03	0	0	0.02	0.05	0.03
PIFL-OSCARPIE	0.03	0	0	0.02	0.01	0.01
IFL-OSCARPIE-A	0.02	0	0	0	0	0.04
SemEval-2022	0.02	0	0	0.01	0.02	0
magpie_random	0.01	0	0	0	0.02	0.01

Table 9 Results for test_f1

Dataset	Gemma3	GPT-OSS	Llama3.1	Mistral	Phi4	Qwen2.5
IFL-C4-A	0.01	0	0	0.02	0.02	0
FLUTE	0.02	0	0	0.01	0.04	0.03
PIFL-OSCARPIE	0.02	0	0	0.01	0.01	0.01
IFL-OSCARPIE-A	0.01	0	0	0	0	0.03
SemEval-2022	0	0	0	0	0.01	0
magpie_random	0.01	0	0	0	0.02	0.01

Table 10 Results for test_sequence_accuracy

Dataset	Gemma3	GPT-OSS	Llama3.1	Mistral	Phi4	Qwen2.5
IFL-C4-A	0.02	0.2	0.2	0.03	0	0.03
FLUTE	0.03	0.09	0.09	0.02	0.03	0.02
PIFL-OSCARPIE	0	0	0	0	0.01	0.01
IFL-OSCARPIE-A	0.06	0.3	0.3	0.03	0	0
SemEval-2022	0.16	0.45	0.44	0.05	0.11	0.18
magpie_random	0	0	0	0	0.01	0

include no labels, incomplete target phrase labeling, incorrect phrase labeling: *shipshape B-Idiom[and] I-Idiom[bristol]* (FLUTE), partial target phrase label combined with non-target phrase label: *get in B-Idiom[touch] I-Idiom[if]* (IFL-C4-A), and over-labeling the phrase with additional words: *B-idiom[lead] I-idiom[the] I-idiom[way] I-idiom[by] I-idiom[trying]* (*P*IFL-OSCAR). One notable exception within the error choices occurs: Mistral correctly predicts *all shipshape* in the FLUTE dataset.

For baseline testing of the slot loss architecture, both BERT and RoBERTa were used. These were finetuned on each dataset with a training/development/test split of 80/10/10 percent, respectively. They were subsequently optimized on the development set. Evaluation occurred on the remaining 10 percent test set. The first positive indication for any model are the test set results compared with the development set results. Evaluation on the test set for each model were consistently higher, while BERT-trained models, rather than RoBERTa-trained models performed better for all metrics, as seen in Table 13.

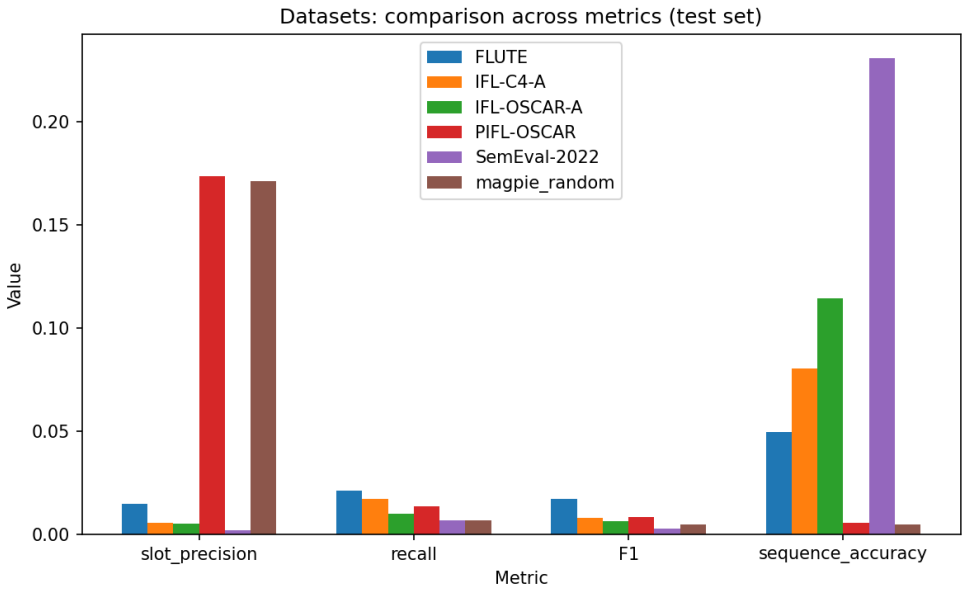

Fig. 6 Average performance of all LLMs by metric and dataset.

When this baseline architecture was applied to the most recent existing dataset, magpie_random, we received competitive results for precision and F1. The results for sequence accuracy achieved state of the art, observed in Fig. 14, compared to the results received by cohesion loss and retranslation [17] at 0.9151.

Each dataset was initially trained and self-evaluated on the same dataset's test split. The models clearly performed best on this setting, but to establish generalizability of each model, a cross evaluation was performed. BERT- and RoBERTa-based models were trained on each dataset, but evaluated on the test sets for the remaining datasets. Development set results were also acquired. Additional results of development data in Figures 9, 10 and tables of raw data are found in the appendix, A. The generalization performance is directly related to the area visualized for sequence accuracy in Fig. 7 and F1 in Fig. 8. The data revealed that the human annotated data did not evaluate well on other datasets, while the models trained on the *P*IFL-OSCAR dataset tested well and generalized the best. According to generalizability, *P*IFL-OSCAR is followed by models trained on magpie_random, IFL-OSCAR-A, IFL-C4-A, with FLUTE and SemEval-2022-trained models generalizing the least.

An interesting subset of sentences within the dataset consists exclusively of idiomatic expressions. The phrase *"Piece of cake!"* exemplifies this category. Such sentences inherently lack surrounding contextual information, resulting in minimal or non-existent intra-sentence cosine similarity data. Consequently, the inclusion of adjacent sentences—both preceding and following—is crucial to capture the inter-sentence cosine similarities that provide broader semantic grounding. By incorporating neighboring context, it becomes possible to more accurately evaluate figurative usage and

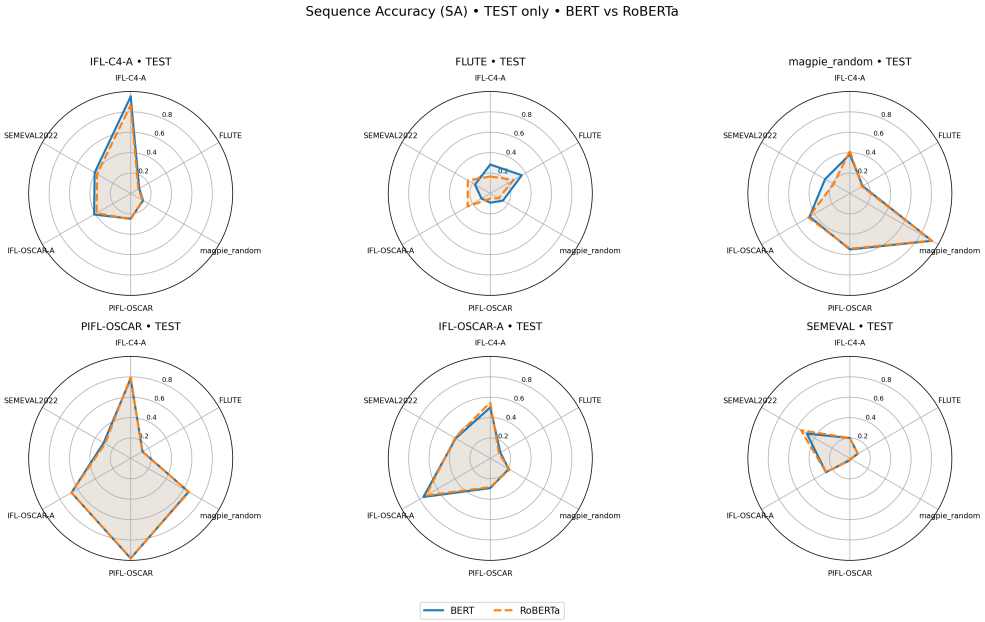

Fig. 7 Cross evaluation results for sequence accuracy on test sets. BERT and RoBERTa models trained with each dataset are evaluated on the test set of all six datasets.

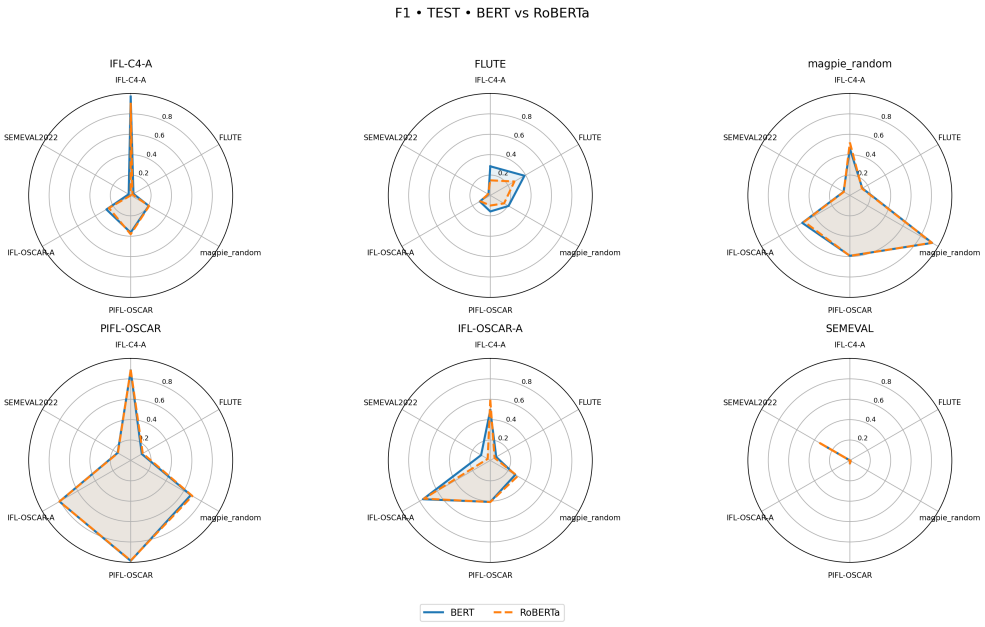

Fig. 8 Cross evaluation results for F1 on test sets. BERT and RoBERTa models trained with each dataset are evaluated on the test set of all six datasets.

Table 11 Examples of errors from chat LLMs for IFL-C4-A, FLUTE, *P*IFL-OSCAR, and IFL-OSCAR-A datasets.

Dataset		Sentence and Model Outputs
IFL-C4-A		
Sentence		Please get in touch if you would like to discuss your grant idea.
Gold Labels		Please B-idiom[get] I-idiom[in] I-idiom[touch] if you would like to discuss your grant idea.
Gemma3	✗	Please B-idiom[get] I-idiom[in] I-idiom[touch] I-idiom[if] I-idiom[you] I-idiom[would] I-idiom[like] to discuss your grant idea.
GPT-OSS	✗	Please get in touch if you would like to discuss your grant idea.
Llama3.1	✗	Please get in touch if you would like to discuss your grant idea.
Mistral	✗	Please get in touch if you would like to discuss your grant idea.
Phi4	✗	Please get in B-idiom[touch] I-idiom[if] you would like to discuss your grant idea.
Qwen2.5	✗	Please B-idiom[get] I-idiom[in] touch if you would like to discuss your grant idea.
FLUTE		
Sentence		All shipshape and bristol fashion.
Gold Labels		B-idiom[All] I-idiom[shipshape] and bristol fashion.
Gemma3	✗	All shipshape and bristol fashion.
GPT-OSS	✗	All shipshape and bristol fashion.
Llama3.1	✗	All shipshape and bristol fashion.
Mistral	✓	B-idiom[All] I-idiom[shipshape] and bristol fashion.
Phi4	✗	All shipshape B-idiom[and] I-idiom[bristol] I-idiom[fashion] .
Qwen2.5	✗	All shipshape B-idiom[and] I-idiom[bristol] fashion.
***P*IFL-OSCAR**		
Sentence		"California can lead the way by trying to bring them home."
Gold Labels		California can B-Idiom[lead] I-idiom[the] I-idiom[way] by trying to bring them home.
Gemma3	✗	B-idiom[California] I-idiom[can] I-idiom[lead] I-idiom[the] I-idiom[way] I-idiom[by] I-idiom[trying] to bring them home.
GPT-OSS	✗	California can lead the way by trying to bring them home.
Llama3.1	✗	California can lead the way by trying to bring them home.
Mistral	✗	California can lead the way by trying to bring them home.
Phi4	✗	California can B-idiom[lead] I-idiom[the] I-idiom[way] I-idiom[by] I-idiom[trying] to bring them home.
Qwen2.5	✗	B-idiom[California] can lead the way by trying to bring them home.
IFL-OSCAR-A		
Sentence		Everyone wants a chance to spill their guts now and then.
Gold Labels		Everyone wants a chance to B-idiom[spill] I-idiom[their] I-idiom[guts] now and then.
Gemma3	✗	B-idiom[Everyone] I-idiom[wants] I-idiom[a] I-idiom[chance] I-idiom[to] I-idiom[spill] I-idiom[their] guts now and then.
GPT-OSS	✗	Everyone wants a chance to spill their guts now and then.
Llama3.1	✗	Everyone wants a chance to spill their guts now and then.
Mistral	✗	Everyone wants a chance to spill their guts now and then.
Phi4	✗	Everyone wants a B-idiom[chance] I-idiom[to] I-idiom[spill] I-idiom[their] I-idiom[guts] now and then.
Qwen2.5	✗	Everyone wants B-idiom[a] I-idiom[chance] to spill their guts now and then.

maintain a coherent representation of semantic relationships, particularly when idioms occur in isolation.

A deeper examination of the errors contributing to these outcomes reveals patterns in the labeling performance, as summarized in Tables, 11, 12, 15, which report BIO tag discrepancies for both the BERT-based models and the chat-oriented large language models (LLMs). During cross-evaluation, a slight reduction was observed between the precision and recall values of models trained on the *P*IFL-OSCAR dataset.

Table 12 Examples of errors from chat LLMs for SemEval-2022 and magpie_random datasets.

Dataset		Sentence and Model Outputs
SemEval 2022		
Sentence		Somehow the Steelers had jitters for this game. And somehow the Steelers laid the biggest goose egg of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
Gold Labels		Somehow the Steelers had jitters for this game. And somehow the Steelers laid the biggest B-Idiom[goose] I-Idiom[egg] of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
Gemma3	✗	B-idiom[Somehow] B-idiom[the] Steelers I-idiom[had] B-idiom[jitters] I-idiom[for] this I-idiom[game]. B-idiom[And] I-idiom[somehow] the I-idiom[Steelers] B-idiom[laid] I-idiom[the] biggest I-idiom[goose] B-idiom[egg] I-idiom[of] a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
GPT-OSS	✗	Somehow the Steelers had jitters for this game. And somehow the Steelers laid the biggest goose egg of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
Llama3.1	✗	Somehow the Steelers had jitters for this game. And somehow the Steelers laid the biggest goose egg of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
Mistral	✗	B-idiom[Somehow] the Steelers I-idiom[had] jitters B-idiom[for] this B-idiom[game]. And I-idiom[somehow] the B-idiom[Steelers] laid the biggest goose egg of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
Phi4	✗	Somehow the Steelers had B-idiom[jitters] I-idiom[for] this game. And somehow the Steelers laid the biggest goose egg of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
Qwen2.5	✗	Somehow B-idiom[the] I-idiom[Steelers] had jitters for this game. And somehow the Steelers laid the biggest goose egg of a game plan of the year. This was as bad as it has ever as been in the Tomlin era.
magpie_random		
Sentence		"Leith , while her voice stayed calm , jumped the gun ."
Gold Labels		Leith , while her voice stayed calm , B-Idiom[jumped] I-Idiom[the] I-Idiom[gun] .
Gemma3	✗	B-idiom[Leith] I-idiom[,] while I-idiom[her] voice I-idiom[stayed] calm B-idiom[,] I-idiom[jumped] the gun .
GPT-OSS	✗	Leith , while her voice stayed calm , jumped the gun .
Llama3.1	✗	Leith , while her voice stayed calm , jumped the gun .
Mistral	✗	B-idiom[Leith] I-idiom[,] while her voice stayed I-idiom[calm] , jumped the gun .
Phi4	✗	Leith , while B-idiom[her] I-idiom[voice] I-idiom[stayed] calm , jumped the gun .
Qwen2.5	✗	Leith , B-idiom[while] I-idiom[her] voice stayed calm , jumped the gun . O O B-idiom I-idiom O O O O O O O O

A widening gap between these metrics can be indicative of over-generalization, a phenomenon influenced by both the architecture of the sequence tagging framework and the methodological emphasis applied during data acquisition. False negatives, as seen in Table 16, also show near and complete misses by the baseline models. Near misses include an *I-Idiom* label in the first position where *B-Idiom* is the logical target at the beginning of a figurative expression. A complete miss includes no label for a labeled word or vice versa.

The annotation of figurative expressions through direct human evaluation was deemed necessary, since figurativeness was inferred when identical or closely matching tokens appeared in the construction of the larger dataset. While the approach facilitated consistent labeling, it also yielded spurious false positives during self-evaluation due to the model's tendency to rely on surface-level lexical features rather than deeper contextual signals. Though, when assessed under cross-evaluation, the generalization

Table 13 Overall result of Bert-based models on our human-annotated datasets.

Dataset	F1	Recall	Sequence Acc.	Slot precision	Model	Batch size
IFL-OSCAR-A	.7708	.8222	.8000	.7255	BERT	4
IFL-OSCAR-A	.6882	.7111	.7286	.6596	RoBERTa	4
*P*IFL-C4-A	.8913	.9318	.8833	.8542	BERT	4
*P*IFL-C4-A	.8542	.9318	.8167	.7885	RoBERTa	4
*P*IFL-OSCAR	.9924	.9947	.9893	.9902	BERT	4
*P*IFL-OSCAR	.9895	.9922	.9857	.9868	RoBERTa	4

Table 14 Overall results of Bert-based models on magpie_random. SOTA achieved for Sequence accuracy in all settings.

Dataset	F1	Recall	Sequence Acc.	Slot precision
BERT test	.9313	.9436	**.9297**	.9193
BERT dev	.9273	.9403	**.9253**	.9147
RoBERTa test	.9289	.9376	**.9254**	.9205
RoBERTa dev	.9270	.9362	**.9226**	.9180

patterns remain stable for *P*IFL-OSCAR, as seen in both sequence accuracy and F1 score Figs. 7,8.

6 Conclusion

A large-scale, targeted dataset of colloquial language was constructed from Creative Commons sources to capture the extensive range of *potential* idiomatic expressions present in natural discourse. Afterwards, a derivative dataset was retrieved from the C4 corpus following equivalent collection and filtering methods. Additional refinement was achieved through human evaluation of both *P*IFL-OSCAR and *P*IFL-C4, resulting in the creation of the IFL-OSCAR-A and IFL-C4-A annotated subsets. These curated datasets collectively contribute to a collection of training-agnostic datasets that explicitly distinguish between idiomatic and non-idiomatic instances of target expressions, and thereby present new opportunities and challenges for consistent evaluation across model architectures and linguistic domains.

Despite these advancements, several limitations merit consideration. The data sanitization process, the presence of possibly machine-generated text in the 2018 snapshot of Common Crawl, and the lexical matching of idiomatic expressions inherently constrain the larger dataset to *possible* rather than *definite* idioms. When combined with cosine similarity measures and linguistic priors—such as syntactic position and part-of-speech distribution—a probabilistic threshold for figurativeness can be estimated and used in hyperparameter optimization.

Future iterations of this dataset could incorporate a confidence metric representing the probability of idiomaticity for each instance. Such an adaptation would facilitate more nuanced model evaluation and training, particularly in determining figurative language type. Extending the dataset to include multilingual variants would enhance

Table 15 Examples of errors from annotated datasets IFL-C4-A and IFL-OSCAR-A.

Errors		
Sentence		It has a smooth , comfortable ride with room for five and generous cargo space **to boot**.
Gold Labels		It has a smooth , comfortable ride with room for five and generous cargo space B-idiom[to] I-idiom[boot].
IFL-C4-A-BERT	✗	It has a smooth , comfortable ride with room for five and generous cargo space to boot.
IFL-C4-A-RoBERTa	✗	It has a smooth , comfortable ride with room for five and generous cargo space to boot.
Sentence		Center parts are great for sleek , **squeaky clean** locks .
Gold Labels		Center parts are great for sleek , B-idiom[squeaky] I-idiom[clean] locks .
IFL-C4-A-BERT	✗	Center parts are great for sleek , B-idiom[squeaky] clean locks .
IFL-C4-A-RoBERTa	✗	Center parts are great for sleek , B-idiom[squeaky] clean locks .
Correct Predictions		
Sentence		How do we seek meaning where none is to be found - - and can we create it **from scratch** ?
Gold Labels		How do we seek meaning where none is to be found - - and can we create it B-Idiom[from] I-Idiom[scratch] ?
IFL-C4-A-BERT	✓	How do we seek meaning where none is to be found - - and can we create it B-Idiom[from] I-Idiom[scratch] ?
IFL-C4-A-RoBERTa	✓	How do we seek meaning where none is to be found - - and can we create it B-Idiom[from] I-Idiom[scratch] ?
Errors		
Sentence		And anyway , isn ' t this a fairly horrible thing to **bank on**
Gold Labels		And anyway , isn ' t this a fairly horrible thing to B-idiom[bank] I-idiom[on] ?
IFL-OSCAR-A-BERT	✗	And anyway , isn ' t this a fairly horrible thing to B-idiom[bank] on ?
IFL-OSCAR-A-RoBERTa	✗	And anyway , isn ' t this a fairly horrible thing to bank on ?
Sentence		Learn more **The bottom line is** we value our customers , and take the additional steps needed in production , research , development and support to provide the finest product and service available .
Gold Labels		Learn more B-idiom[The] I-idiom[bottom] I-idiom[line] I-idiom[is] we value our customers , and take the additional steps needed in production , research , development and support to provide the finest product and service available .
IFL-OSCAR-A-BERT	✗	Learn more I-idiom[The] I-idiom[bottom] I-idiom[line] I-idiom[is] we value our customers , and take the additional steps needed in production , research , development and support to provide the finest product and service available .
IFL-OSCAR-A-RoBERTa	✗	Learn more B-idiom[The] I-idiom[bottom] I-idiom[line] is we value our customers , and take the additional steps needed in production , research , development and support to provide the finest product and service available .
Correct Predictions		
Sentence		This is a smaller office , so the wait times are **next to nothing** !
Gold Labels		This is a smaller office , so the wait times are B-idiom[next] I-idiom[to] I-idiom[nothing] !
IFL-OSCAR-A-BERT	✓	This is a smaller office , so the wait times are B-idiom[next] I-idiom[to] I-idiom[nothing] !
IFL-OSCAR-A-RoBERTa	✓	This is a smaller office , so the wait times are B-idiom[next] I-idiom[to] I-idiom[nothing] !

Table 16 False negatives observed in error analysis. The target sentence contained other figurative language, extended figurative expression, or the model falsely predicted without a beginning idiom word.

FALSE NEGATIVES	
Gold Labels	When Aurora first came to London in 2016 she never thought that the songs she had collected over the years would ever see I-idiom[the] I-idiom[light] of day.
IFL-OSCAR-A-BERT	When Aurora first came to London in 2016 she never thought that the songs she had collected over the years would ever see I-idiom[the] I-idiom[light] I-idiom[of] I-idiom[day].
IFL-OSCAR-A-RoBERTa	When Aurora first came to London in 2016 she never thought that the songs she had collected over the years would ever see I-idiom[the] I-idiom[light] I-idiom[of] I-idiom[day].
Gold Labels	company name On the same note, platform fees, paid trading indicators and other trading software can also be an issue and you should evaluate your need for the trading platform that has all the B-idiom[bells] I-idiom[and] I-idiom[whistles].
IFL-OSCAR-A-BERT	company name B-idiom[On] I-idiom[the] I-idiom[same] I-idiom[note], platform fees, paid trading indicators and other trading software can also be an issue and you should evaluate your need for the trading platform that has all the B-idiom[bells] I-idiom[and] I-idiom[whistles].
IFL-OSCAR-A-RoBERTa	Correct according to gold.
Gold Labels	[username]'s instructions are not only in Dutch, but pretty bare bones, so I didn't bother translating them.
IFL-OSCAR-A-BERT	(no prediction)
IFL-OSCAR-A-RoBERTa	[username]'s instructions are not only in Dutch, but pretty B-idiom[bare] I-idiom[bones], so I didn't bother translating them.

its generalizability across languages and enable broader cross-linguistic analyses of idiomatic expressions within natural language understanding frameworks.

A Tables of Raw Cross Evaluation Values

Table 17 Cross evaluation results on BERT-based models (Test set). For each column, each metric is shown. (Sequence Accuracy/F1/Precision/Recall)

Dataset	IFL-C4-A	FLUTE	magpie-random	PIFL-OSCAR	IFL-OSCAR-A	SemEval-2022
IFL-C4-A-BERT	95/96.91/95/97.2	9.68/3.12/13.95/1.75	14.02/20.44/33.2/14.76	24.76/36.46/57.83/26.62	41.43/27.5/36.67/22	40.81/2.29/5.76/1.43
IFL-C4-A-RoBERTa	86.67/90.91/88.24/93.75	8.62/1.99/6.67/1.17	13.44/20.78/38.23/14.27	25.42/38.08/65.08/27.77	38.57/24.66/39.13/18	38.06/0/0/0
FLUTE-BERT	28.33/28.57/26.32/31.25	35.68/38.7/37.55/39.91	14.24/20.67/19.22/22.35	9.06/15.74/14.89/16.69	10/11.85/9.41/16	17.32/2.06/1.55/3.09
FLUTE-RoBERTa	16.67/14.81/18.18/12.5	26.53/27.31/28.55/26.17	9.46/15.62/20.75/12.53	5.08/9.92/13.14/7.97	25.71/11.36/13.16/10	25.33/2.49/2.39/2.62
magpie-random-BERT	38.33/45.99/44.23/47.92	14.46/14.07/14.41/13.74	92.97/93.13/91.93/94.36	54.95/59.21/61.89/56.76	45.71/53.91/47.69/62	28.08/7.14/7.71/6.65
magpie-random-RoBERTa	41.67/52/50/54.17	13.66/13.64/13.83/13.45	92.54/92.89/92.05/93.76	54.18/59.31/63.21/55.87	47.14/51.67/44.29/62	18.37/6.89/5.88/8.33
PIFL-OSCAR-BERT	78.33/88.07/78.69/100	13.39/12.56/13.56/11.69	65.54/68.07/66.93/69.26	98.06/98.55/98.05/99.04	67.14/80.65/67.57/100	30.58/14.79/15.09/14.49
PIFL-OSCAR-RoBERTa	80/88.89/80/100	13.13/13.67/13.89/13.45	66.03/69.75/70.63/68.89	97.71/98.19/97.79/98.59	67.14/80.65/67.57/100	29/14.47/14.42/14.52
IFL-OSCAR-A-BERT	20/0/0/0	9.02/0/0/0	1.01/1.03/10.21/0.54	1.92/3.11/22.89/1.67	27.14/0/0/0	48.68/27.27/35.63/22.09
IFL-OSCAR-A-RoBERTa	20/0/0/0	9.15/0/0/0	0.92/0.94/9.05/0.49	1.98/3.04/19.94/1.65	27.14/0/0/0	55.25/35.57/45.86/29.05
SEMEVAL-BERT	50/49.38/60.61/41.67	11.41/6.73/18.92/4.09	20.92/28.28/38.57/22.33	29.13/40.72/54.76/32.41	75.71/76.19/72.73/80	39.24/10.39/17.61/7.36
SEMEVAL-RoBERTa	55/59.46/84.62/45.83	9.81/4.86/11.6/3.07	21.88/30.95/47.59/22.93	28.05/40.85/59.52/31.09	72.86/75/72.22/78	39.89/2.96/6.61/1.9

Table 18 Cross evaluation results on BERT-based models (Dev set). For each column, each metric is shown. (Sequence Accuracy/F1/Precision/Recall)

Dataset	IFL-C4-A	FLUTE	magpie-random	PIFL-OSCAR	IFL-OSCAR-A	SemEval-2022
IFL-C4-A-BERT	88.33/89.13/85.42/93.18	11.55/3.46/17.33/1.92	14.584/21.17/33.81/15.41	24.38/36.03/56.03/26.55	61.43/53.66/59.46/48.89	45.87/0/0/0
IFL-C4-A-RoBERTa	81.67/85.42/78.85/93.18	11.95/4.76/15.57/2.81	14.63/22.39/41.07/15.39	24.91/37.42/59.85/27.22	50/34.86/48/26.67	43.57/0.41/0.81/0.28
FLUTE-BERT	26.67/23.3/20.34/27.27	38.38/38.98/37.41/40.68	13.66/20.79/19.16/22.74	9.39/16.17/15.42/17.01	21.43/23.33/18.67/31.11	19.08/1.39/1.01/2.21
FLUTE-RoBERTa	26.67/17.58/17.02/18.18	29.75/31.71/32.21/31.21	9.36/15.24/19.85/12.37	5.71/10.49/13.97/8.41	30/16.67/15.69/17.78	27.19/0.48/0.43/0.55
magpie-random-BERT	36.67/46.46/41.82/52.27	14.74/14.99/15.19/14.79	92.53/92.73/91.47/94.03	55.46/59.48/61.85/57.29	54.29/65.49/54.41/82.22	31.53/6.49/6.65/6.35
magpie-random-RoBERTa	41.67/53.47/47.37/61.36	14.08/13.74/14.19/13.31	92.26/92.7/91.8/93.62	54.24/58.83/61.94/56.02	48.57/61.95/51.47/77.78	18.67/5.35/4.36/6.91
PIFL-OSCAR-BERT	68.33/82.24/69.84/100	15.14/14.84/15.91/13.91	66.03/68.34/67.19/69.54	98.34/98.82/98.31/99.35	61.43/76.27/61.64/100	33.56/11.59/11.32/11.88
PIFL-OSCAR-RoBERTa	68.33/82.24/69.84/100	14.21/13.85/13.65/14.05	66.77/70.49/70.82/70.17	97.88/98.45/97.93/98.98	61.43/76.52/62.86/97.78	28.82/8.39/8/8.84
IFL-OSCAR-A-BERT	26.67/0/0/0	10.22/0/0/0	0.88/0.94/10.05/0.49	1.75/2.89/21.79/1.55	34.29/0/0/0	53.72/26.77/32.17/22.93
IFL-OSCAR-A-RoBERTa	28.33/0/0/0	10.09/0/0/0	0.76/0.68/6.27/0.36	1.79/2.95/20.16/1.59	35.71/0/0/0	55.75/33.67/42.44/27.9
SEMEVAL-BERT	51.67/44.16/51.52/38.64	11.69/5.57/15.33/3.4	21.69/28.71/38.29/22.96	29.67/41.61/55.22/33.38	80/77.08/72.55/82.22	45.06/5.55/12.26/3.59
SEMEVAL-RoBERTa	63.33/66.67/80.65/56.82	12.35/6.88/15.31/4.44	22.01/31.76/47.74/23.79	28.22/40.91/59.42/31.19	72.86/68.82/66.67/71.11	43.44/1.26/2.65/0.83

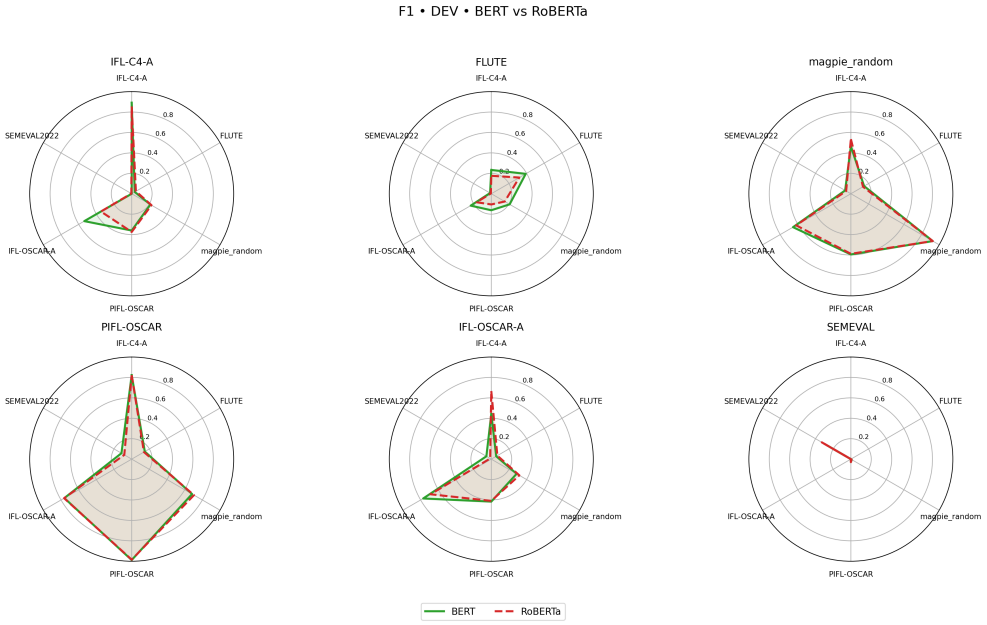

Fig. 9 Cross evaluation results for F1 on test sets. BERT and RoBERTa models trained with each dataset are evaluated on the dev set of all six datasets.

References

[1] Ortiz Suárez, P.J., Sagot, B., Romary, L.: Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In: Bański, P., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., Kupietz, M., Lüngen, H., Iliadi, C. (eds.) Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9–16. Leibniz-Institut für Deutsche Sprache, Mannheim (2019). https://doi.org/10.14618/ids-pub-9021 . https://doi.org/10.14618/ids-pub-9021

[2] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (2023). https://arxiv.org/abs/1910.10683

[3] Moussallem, D., Sherif, M.A., Esteves, D., Zampieri, M., Ngonga Ngomo, A.-C.: Lidioms: A multilingual linked idioms data set. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), ??? (2018)

[4] Saxena, P., Paul, S.: EPIE Dataset: A Corpus For Possible Idiomatic Expressions (2020). https://arxiv.org/abs/2006.09479

[5] Haagsma, H., Bos, J., Nissim, M.: MAGPIE: A large corpus of potentially

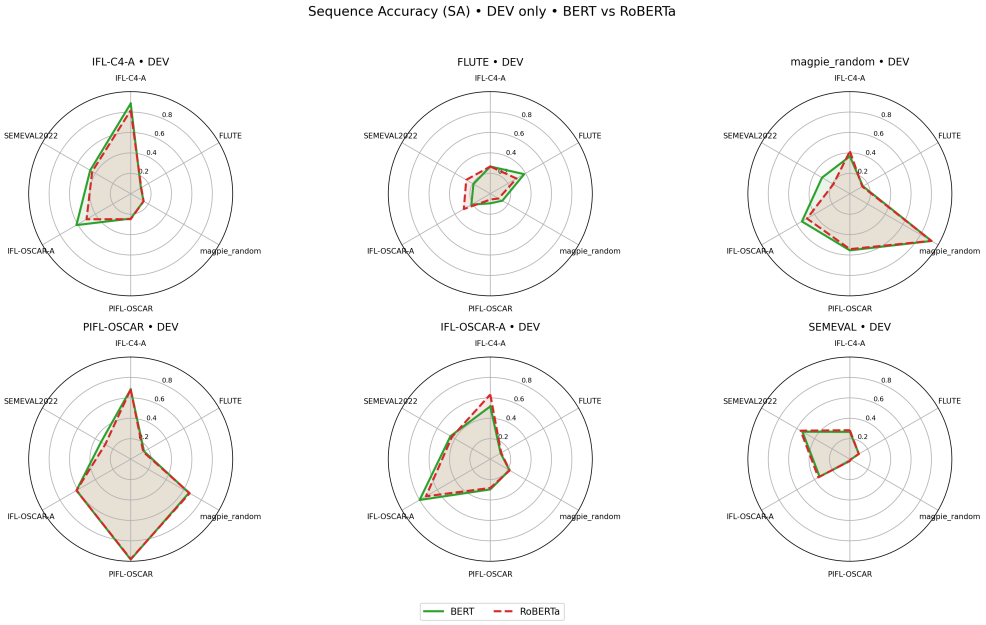

Fig. 10 Cross evaluation results for sequence accuracy on test sets. BERT and RoBERTa models trained with each dataset are evaluated on the dev set of all six datasets.

idiomatic expressions. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 279–287. European Language Resources Association, Marseille, France (2020). https://aclanthology.org/2020.lrec-1.35/

[6] Chakrabarty, T., Saakyan, A., Ghosh, D., Muresan, S.: FLUTE: Figurative language understanding through textual explanations. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 7139–7159. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.481 . https://aclanthology.org/2022.emnlp-main.481/

[7] Madabushi, H.T., Gow-Smith, E., Garcia, M., Scarton, C., Idiart, M., Villavicencio, A.: Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 107–121. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.semeval-1.13 . https://aclanthology.org/2022.semeval-1.13/

[8] Sinclair, J.M.: Corpus, Concordance, Collocation. Oxford University Press, Oxford, UK (1991)

[9] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423

[10] Wang, S., Thompson, L., Iyyer, M.: Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration (2021). https://arxiv.org/pdf/2109.06304

[11] Qin, R., Luo, H., Fan, Z., Ren, Z.: IBERT: Idiom Cloze-style reading comprehension with Attention (2021). https://arxiv.org/abs/2112.02994

[12] Zhou, J., Zeng, Z., Gong, H., Bhat, S.: Idiomatic expression paraphrasing without strong supervision. In: AAAI-22 Technical Tracks 10. Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, pp. 11774–11782. Association for the Advancement of Artificial Intelligence, ??? (2022). https://doi.org/10.1609/aaai.v36i10.21433 . Publisher Copyright: Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 36th AAAI Conference on Artificial Intelligence, AAAI 2022 ; Conference date: 22-02-2022 Through 01-03-2022

[13] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202 . https://aclanthology.org/N18-1202/

[14] Škvorc, T., Gantar, P., Robnik-Šikonja, M.: Mice: Mining idioms with contextual embeddings. Knowledge-Based Systems **235**, 107606 (2022) https://doi.org/10.1016/j.knosys.2021.107606

[15] Kurfalı, M., Östling, R.: Disambiguation of potentially idiomatic expressions with contextual embeddings. In: Markantonatou, S., McCrae, J., Mitrović, J., Tiberius, C., Ramisch, C., Vaidya, A., Osenova, P., Savary, A. (eds.) Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, pp. 85–94. Association for Computational Linguistics, online (2020). https://aclanthology.org/2020.mwe-1.11/

[16] Takahashi, R., Sasano, R., Takeda, K.: Leveraging three types of embeddings from masked language models in idiom token classification. In: Nastase, V., Pavlick, E., Pilehvar, M.T., Camacho-Collados, J., Raganato, A. (eds.) Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pp. 234–239. Association for Computational Linguistics, Seattle, Washington (2022). https://

doi.org/10.18653/v1/2022.starsem-1.21 . https://aclanthology.org/2022.starsem-1.21/

[17] Yayavaram, A., Yayavaram, S., Upadhyay, P.D., Das, A.: BERT-based idiom identification using language translation and word cohesion. In: Bhatia, A., Bouma, G., Doğruöz, A.S., Evang, K., Garcia, M., Giouli, V., Han, L., Nivre, J., Rademaker, A. (eds.) Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) at LREC-COLING 2024, pp. 220–230. ELRA and ICCL, Torino, Italia (2024). https://aclanthology.org/2024.mwe-1.26/

[18] Phelps, D., Pickard, T., Mi, M., Gow-Smith, E., Villavicencio, A.: Sign of the times: Evaluating the use of large language models for idiomaticity detection. In: Bhatia, A., Bouma, G., Doğruöz, A.S., Evang, K., Garcia, M., Giouli, V., Han, L., Nivre, J., Rademaker, A. (eds.) Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, pp. 178–187. ELRA and ICCL, Torino, Italia (2024). https://aclanthology.org/2024.mwe-1.22/

[19] De Luca Fornaciari, F., Altuna, B., Gonzalez-Dios, I., Melero, M.: A hard nut to crack: Idiom detection with conversational large language models. In: Ghosh, D., Muresan, S., Feldman, A., Chakrabarty, T., Liu, E. (eds.) Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), pp. 35–44. Association for Computational Linguistics, Mexico City, Mexico (Hybrid) (2024). https://doi.org/10.18653/v1/2024.figlang-1.5 . https://aclanthology.org/2024.figlang-1.5/

[20] Donthi, S., Spencer, M., Patel, O.B., Doh, J.Y., Rodan, E., Zhu, K., O'Brien, S.: Improving LLM abilities in idiomatic translation. In: Hettiarachchi, H., Ranasinghe, T., Rayson, P., Mitkov, R., Gaber, M., Premasiri, D., Tan, F.A., Uyangodage, L. (eds.) Proceedings of the First Workshop on Language Models for Low-Resource Languages, pp. 175–181. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2025). https://aclanthology.org/2025.loreslm-1.13/

[21] He, W., Idiart, M., Scarton, C., Villavicencio, A.: Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Findings of the ACL: ACL 2024, pp. 12473–12485. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.findings-acl.741 . https://aclanthology.org/2024.findings-acl.741/

[22] Zeng, Z., Bhat, S.: Idiomatic expression identification using semantic compatibility. Transactions of the Association for Computational Linguistics **9**, 1546–1562 (2021) https://doi.org/10.1162/tacl_a_00442

[23] Tayyar Madabushi, H., Gow-Smith, E., Scarton, C., Villavicencio, A.: AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in

pre-trained language models. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3464–3477. Association for Computational Linguistics, Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.294 . https://aclanthology.org/2021.findings-emnlp.294/

[24] Cook, P., Fazly, A., Stevenson, S.: The vnc-tokens dataset. In: Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008), pp. 19–22. European Language Resources Association (ELRA), Marrakech, Morocco (2008). https://www.cs.toronto.edu/~suzanne/papers/CFS2008.pdf

[25] Miletić, F., Walde, S.: Semantics of multiword expressions in transformer-based models: A survey. Transactions of the Association for Computational Linguistics **12**, 593–612 (2024) https://doi.org/10.1162/tacl_a_00657

[26] Zeng, Z., Bhat, S.: Idiomatic expression identification using semantic compatibility. Transactions of the Association for Computational Linguistics **9**, 1546–1562 (2021) https://doi.org/10.1162/tacl_a_00442

[27] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks (2014). https://arxiv.org/abs/1409.3215

[28] Phuong, N.M., Le, T., Minh, N.L.: Cae: Mechanism to diminish the class imbalanced in slu slot filling task. In: Bădică, C., Treur, J., Benslimane, D., Hnatkowska, B., Krótkiewicz, M. (eds.) Advances in Computational Collective Intelligence. Communications in Computer and Information Science, vol. 1653, pp. 150–163. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16210-7_12

[29] Chen, Q., Zhuo, Z., Wang, W.: BERT for Joint Intent Classification and Slot Filling (2019). https://doi.org/10.48550/arXiv.1902.10909 . https://arxiv.org/abs/1902.10909

[30] Meta Platforms, I.: Llama 3.1: Model Card. https://github.com/meta-llama/llama3. Release date: July 23 2024 (2024)

[31] Jiang, A., Sablayrolles, A., Mensch, C.e.a.: Mistral 7b: A 7-billion-parameter language model. arXiv preprint arXiv:2310.06825 (2023)

[32] DeepMind), G.T.G.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)

[33] Research, M.: Phi-4 technical report. arXiv preprint arXiv:2412.08905 (2024)

[34] Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M.,

Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 Technical Report (2025). https://arxiv.org/abs/2412.15115