Using LLMs to Advance Idiom Corpus Construction

Doğukan Arslan[*] and **Hüseyin Anıl Çakmak**[*] and **Gülşen Eryiğit**[*] and **Joakim Nivre**[†]
[*]ITU NLP Research Group, Istanbul Technical University, Türkiye;
[†]Department of Linguistics and Philology, Uppsala University, Sweden
[*]{arslan.dogukan, cakmakh19, gulsen.cebiroglu}@itu.edu.tr
[†]joakim.nivre@lingfil.uu.se

Abstract

Idiom corpora typically include both idiomatic and literal examples of potentially idiomatic expressions, but creating such corpora traditionally requires substantial expert effort and cost. In this article, we explore the use of large language models (LLMs) to generate synthetic idiom corpora as a more time- and cost-efficient alternative. We evaluate the effectiveness of synthetic data in training task-specific models and testing GPT-4 in few-shot prompting setting using synthetic data for idiomaticity detection. Our findings reveal that although models trained on synthetic data perform worse than those trained on human-generated data, synthetic data generation offers considerable advantages in terms of cost and time. Specifically, task-specific idiomaticity detection models trained on synthetic data outperform the general-purpose LLM that generated the data when evaluated in a zero-shot setting, achieving an average improvement of 11 percentage points across four languages. Moreover, synthetic data enhances the LLM's performance, enabling it to match the task-specific models trained with synthetic data when few-shot prompting is applied.

1 Introduction

An idiom is a linguistic expression the meaning of which cannot be derived compositionally from the literal meaning of its parts. For example, the English idiom *break a leg* is used to wish someone good luck, rather than being taken literally as an instruction to cause physical harm. Due to this unique nature, idioms can negatively impact the performance of models in various tasks, such as machine translation, word-sense disambiguation, and information retrieval (Korkontzelos et al., 2013; Isabelle et al., 2017).

Idiom corpora are essential for enhancing performance in numerous tasks, as they provide both idiomatic and non-idiomatic examples to help models better differentiate between literal and figurative meanings. Training models on a diverse and well-structured idiom corpus can reduce problems such as incorrect translations (Fadaee et al., 2018) or misinterpretation of idiomatic expressions (Adewumi et al., 2022). Moreover, idioms present a significant challenge for language learners, who often struggle with the non-literal meanings and cultural nuances embedded in these expressions (Cieślicka, 2015). Comprehensive idiom corpora can support the development of educational resources and tools designed to help learners master idiomatic usage more effectively. Consequently, both computers and humans require high-quality samples that exemplify idiom usage scenarios and patterns.

Traditional approaches to constructing idiom corpora, such as those relying on the annotation of natural text (Cook et al., 2008), face several challenges. These include unbalanced distributions of idiomatic versus non-idiomatic examples, a lack of diversity in surface forms, and issues related to data scarcity. While recent methods, such as obtaining idiomatic sentences from native speakers via gamified crowdsourcing platforms (Eryiğit et al., 2022), offer potential solutions, they still have notable limitations and continue to be time-consuming and costly, as they require the involvement of native speakers for effective execution. Due to the challenging nature of the data collection process, only a handful of studies have presented idiom corpora that include both idiomatic and non-idiomatic examples. These corpora are mostly limited to a few languages and a small set of idioms (see Table 1).

Recently, large language models (LLMs), such as GPT-3 (Brown et al., 2020), have shown their effectiveness as generators in few-shot (Wang et al., 2021) and zero-shot (Gao et al., 2023) settings, and have been utilized to generate training data for downstream tasks (Meng et al., 2022). In this article, we use GPT-4 to generate idiomatic in-

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 21–31

Dataset	#Sentences	#Idioms	Language
VNC-Tokens* (Cook et al., 2008)	2,566	53	en
Open-MWE* (Hashimoto and Kawahara, 2009)	102,856	146	ja
Sporleder and Li (Sporleder and Li, 2009)	3,964	17	en
IDIX (Sporleder et al., 2010)	5,836	78	en
SemEval-2013 Task 5b (Korkontzelos et al., 2013)	4,350	65	en
PARSEME (Savary et al., 2015)	274,376	13,755	bg, cs, fr, de, he, it, lt, mt, el, pl, pt, ro, sl, es, sv, tr
MAGPIE (Haagsma et al., 2020)	56,622	2,007	en
EPIE (Saxena and Paul, 2020)	25,206	717	en
AStitchInLanguageModels	6,430	336	en, pt
ID10M$_{silver}$ (Tedeschi et al., 2022)	800	470	de, en, es, it
ID10M$_{gold}$ (Tedeschi et al., 2022)	262,781	10,118	de, en, es, fr, it, ja, nl, pl, pt, zh
SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022)	8,683	50	en, gl, pt
Dodiom* (Eryiğit et al., 2022)	12,706	73	it, tr

Table 1: Overview of various idiom corpora, listing the number of sentences, idioms, and the languages they cover (based on ISO 639-1 language codes). Datasets used in this article are marked with an asterisk.

stances, providing a time and cost-efficient alternative to human-involved methods. We generate sentence examples containing idioms in English, Italian, Turkish, and Japanese using zero-shot and enhanced prompting settings. To assess the quality of the LLM-produced corpora against human-generated data, we fine-tune relatively smaller models (i.e., BERT variants) specifically for the task of idiomaticity detection. Models fine-tuned on synthetic data never reach the performance of those trained on human-generated data, likely due to LLMs' potential struggle to generate data instances that fully capture real-world scenarios. However, the results show that with further refinement in the reasoning process of LLMs for synthetic data generation and the usage of synthetic data in few-shot prompting settings, LLM-generated synthetic data could yield more competitive outcomes, highlighting potential for future development. Notably, task-specific models trained on synthetic data outperformed the large language model that generated it (in zero-shot setting) when tested on human datasets, demonstrating the effectiveness of leveraging large models for data generation and then training smaller models and offers a more efficient and scalable approach to model development while also indicating the potential for LLMs to perform better after fine-tuning.

We also investigate the effect of prompt engineering on dataset quality by comparing zero-shot and enhanced prompting through separate model training. Zero-shot prompting yields slightly higher quality data,[1] likely due to the enhanced prompt's complexity. Additionally, we train multilingual BERT models using the constructed data sets for all five languages (English, Italian, Japanese, Turkish). The results show minimal performance differences, suggesting that synthetic data can effectively train multilingual models without significant loss compared to monolingual models.

In summary, our main contributions can be listed as:

1. We construct synthetic idiom corpora for English, Japanese, Italian and Turkish using GPT-4.
2. We investigate the impact of synthetic datasets on the idiomaticity detection task.
3. We examine the impact of prompt style on creating synthetic idiom data.
4. We investigate the performance of different task-specific BERT models and GPT-4 on the idiomaticity detection task.
5. We investigate the effect of few-shot prompting on GPT-4's performance in the idiomaticity detection task.

[1] Here, quality refers to the data's ability to improve model performance in the idiomaticity detection task.

6. We investigate the impact of multilingual training on the idiomaticity detection task.

The constructed corpora, along with the code for synthetic data generation and training and testing models for idiomaticity detection, are available on GitHub.[2]

2 Background and Related Work

Idiom corpora are corpora that include sentences containing potentially idiomatic expressions (PIEs), where these expressions are used in both idiomatic and literal senses in different contexts. The process of constructing an idiom corpus generally involves three steps: (1) selecting a list of idioms from phrases identified in previous studies (Hashimoto and Kawahara, 2009; Tayyar Madabushi et al., 2021) or from dictionaries (Sporleder and Li, 2009; Haagsma et al., 2020), with optional filtering based on certain rules (Saxena and Paul, 2020), frequency (Sporleder et al., 2010), or expert judgment (Cook et al., 2008); (2) obtaining sentences that contain PIEs from existing corpora (Sporleder et al., 2010), the web (Tayyar Madabushi et al., 2022), or directly from native speakers (Eryiğit et al., 2022); and (3) labeling the sentences based on the usage sense, typically as idiomatic or literal, using native speakers or language experts (Tedeschi et al., 2022). In Table 1, we provide an overview of various idiom corpora, listing the number of sentences, idioms, and the languages they cover.

Synthetic data generation involves creating artificial datasets that mimic the statistical properties and patterns of real-world data. Recently, LLMs have emerged as powerful tools for generating synthetic data, leveraging their vast training on diverse textual data to produce high-quality, contextually relevant examples (Long et al., 2024). The general paradigms for synthetic data generation with LLMs typically involve prompt engineering, where carefully designed prompts guide the model to produce desired outputs, and iterative refinement, where generated data is evaluated and adjusted for quality and relevance. For instance, Li et al. (2023) utilizes LLMs to generate synthetic data for classification tasks, and analyzed the effect of task and instance subjectivity on model performance, finding a negative impact. Tang et al. (2023) demonstrates that directly utilizing LLMs for tasks like clinical text mining may result in poor performance and raise privacy issues related to patient information; however, creating high-quality synthetic labeled data with LLMs and subsequently fine-tuning a smaller model can substantially improve the performance of downstream tasks. Additionally, Heng et al. (2024) introduces a cost-efficient strategy to leverage LLMs with moderate NER capabilities for generating high-quality NER datasets, which significantly improves performance compared to traditional data generation methods.

3 Methodology

To construct the idiom corpora presented in this article, we select a list of PIEs identified in previous research that provides a diverse set of idiomatic expressions in different languages. Specifically, we choose the PIEs identified by Cook et al. (2008) for English, Hashimoto and Kawahara (2009) for Japanese, and Eryiğit et al. (2022) for Italian and Turkish.

For synthetic data generation, we prompt GPT-4 (specifically gpt-4-0125-preview) to generate a sentence containing an idiomatic or literal use of an identified PIE in two settings: zero-shot prompting and enhanced prompting (See Appendix A, Figure 1 and Figure 2). In both settings, the prompts are always given in the target language and the system prompt instructs the model to generate sentences as if it is proficient in the target language, using it in rich and creative ways. The model is specifically asked to retain the lemma of the idiom constituents, since syntactic operations for idioms are mainly restricted by the idiom's individual components and its overall idiomatic meaning (Cacciari and Tabossi, 2014). Additionally, it is prompted to avoid the use of human names, as our prior prompting trials indicate that including names results in poor-quality samples. In the enhanced prompting setting, the model is further instructed to avoid repeating previously generated sentences, and it is observed that explicitly encouraging creativity (e.g., prompting the model to be creative) sometimes results in similar sentence structures.

In the zero-shot setting, the model is introduced to a PIE and simply asked to generate sentences using it. In the enhanced setting, a two-stage data generation approach is applied using the chain-of-thought method (Wei et al., 2024). First, the model is presented with a PIE and its use cases, and then it is asked to generate use cases for another target PIE. In the second step, the model is

[2] github.com/itunlplab/idiom-corpus-llm

instructed to generate sentences based on those use cases, incorporating diverse grammatical structures, including declarative-interrogative forms, affirmative-negative constructions, variations in sentence length, and inserting additional words between the components of the idiom. This approach aims to ensure diversity in sentence structures within the generated corpus. Additionally, the model is encouraged once in a while to generate sentences "as-if it is a human" and to be "creative", to prevent it from simply paraphrasing previous answers. Illustrations of the zero-shot and enhanced prompting settings are provided in Figure 1 and Figure 2, respectively, which can be found in Appendix A.

For each PIE in the aforementioned corpora, we generate 200 sentences using GPT-4 through the OpenAI API, with each PIE appearing in both its idiomatic and literal senses, equally represented with 100 sentences for each sense. Of these, 60 sentences are generated using zero-shot prompting, while 40 are generated using enhanced prompting. The average cost of generating each sentence is approximately $0.004. The overall statistics for the generated datasets are summarized in Table 2.

Language	#Idioms	#Sentences
English	53	10,600
Japanese	47	9,400
Italian	37	7,400
Turkish	36	7,200

Table 2: An overview of the generated datasets, including the number of idioms used and the generated sentences for each language.

4 Experiments

To evaluate the quality of the synthetically generated datasets, we applied it to n-shot prompting of GPT-4 and fine-tuning smaller models specifically for the task of idiomaticity detection. Additionally, to examine the effects of different prompting techniques on data generation, we fine-tune separate models using examples obtained from zero-shot prompting and enhanced prompting, allowing for a comparison between these two approaches. Finally, we measure and compare the performance of multilingual models fine-tuned on idioms from multiple languages against monolingual models to assess the impact of multilingual idiom inclusion.

4.1 N-Shot Prompting

We evaluate GPT-4's performance in idiomaticity detection across various n-shot prompting settings, including zero-shot, one-shot, and few-shot scenarios, using both synthetic and human-generated data. In the zero-shot setting, GPT-4 is prompted to determine whether a given sentence contains a PIE used in a figurative or literal sense, with the expected output being 1 or 0, respectively. For the one-shot setting, GPT-4 is provided with two example sentences—one illustrating a figurative usage of a PIE and the other illustrating a literal usage. We conduct experiments where example sentences containing PIEs are either randomly selected or include the same PIE as the test sentence.

To investigate the impact of the number of sample sentences, we extend the experiments to include 3 and 5 synthetically generated examples for figurative and literal senses, all containing the same PIE as the test sentence. Additionally, we examine how the order of example presentation influences GPT-4's performance by presenting literal sentence examples before figurative ones.

Further experiments incorporate human-generated data into the prompts. Since some idioms in the English dataset exhibit only figurative or only literal meanings, missing examples are substituted with randomly selected entries from the dataset. This strategy is intended to simulate real-world scenarios more accurately by addressing gaps in the dataset.

4.2 Task-Specific Fine-tuning

To determine whether the generated datasets are sufficiently comprehensive and of comparable quality to human-produced data, we fine-tune various BERT variants such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), DistilBERT (Sanh et al., 2020), and language-specific BERTS such as Japanese BERT,[3] Italian BERT (Schweter, 2020b), and BERTurk (Schweter, 2020a) on the task of idiomaticity detection using synthetically generated datasets. This task involves identifying whether the PIE in a given sentence is used figuratively or literally, and classifying the sentences accordingly. Classification is performed using a linear layer added on top of the models. This layer takes the hidden state of the [CLS] token as input and outputs a vector of size equal to the number of target classes. The models are fine-tuned

[3] github.com/cl-tohoku/bert-japanese

with a batch size of 8 and a learning rate of 5e-6 for 4 epochs. Experiments are repeated three times using different seed values (5, 42, 1773).

During the training phase, 80% of the synthetic datasets are used for training, and 20% for validation, maintaining the zero-shot to enhanced prompting ratio of 60% to 40% in both sets. For comparison, the same models are also trained on human-produced data, utilizing 60% of each dataset for training and 10% for validation. The relative sizes of the synthetic and human-produced datasets vary depending on the language. For English and Japanese, the sizes are highly unbalanced in favor of synthetic or human-produced data, respectively. In contrast, for Italian and Turkish, the synthetic and human-produced datasets are closer in size, but synthetic data remains slightly larger. In the test phase, 30% of the real-world datasets are employed to evaluate both types of models, ensuring a consistent comparison between those trained with synthetic and human-produced data. The test set sizes vary considerably, with Japanese having a much larger test set (15,239 examples) compared to English (807 examples), Italian (2,284 examples), and Turkish (2,084 examples).

In the English dataset, sentences labeled as "unknown"[4] are excluded from both the training and testing sets. In the Japanese dataset, to align with the other datasets, idioms containing over 900 samples were selected, resulting in a focus on 47 idioms for further analysis instead of using the original dataset, which consists of 146 idioms. All examples from the Italian and Turkish datasets are used directly without any filtering. For idioms in the human-generated datasets, if only one sentence represented the idiom, it is included in the training set. If there are two sentences, they are distributed between the training and validation sets. For idioms with at least three sentences, the sentences are distributed across the sets based on the above ratios, ensuring that at least one sentence appeared in each dataset.

4.3 Zero-Shot vs. Enhanced Prompting

To investigate the contributions of the two distinct prompting methods used for producing synthetic data, the previously mentioned models are trained separately on the data generated by each prompting approach (i.e., zero-shot and enhanced prompting) and tested with the human-generated data as in the earlier step. To ensure a fair comparison between the two prompting strategies and to address potential concerns related to data imbalance, we also conduct tests using 40 sample subsets for the zero-shot prompting (i.e., zero-shot filtered). Additionally, to investigate and compare the diversity of human-generated data and synthetic data constructed using zero-shot prompting and enhanced prompting, we apply the remote clique score (the average mean distance of a data instance to other instances) and the Chamfer distance score (the average minimum distance of a data instance to other instances).

4.4 Multilingual Idiomaticity Detection

To assess the impact of the training set on model performance in the idiomaticity detection task, we train models using synthetic data, combining all languages into a single multilingual training set. Specifically, we merge all available languages, allocating 80% of the data for training and 20% for validation. The trained multilingual models are then evaluated on human-generated test sets, following the same procedure as in previous steps.

5 Results

This section summarizes the findings from our experiments, which include comparing the performance of n-shot prompted GPT-4 and BERT variants fine-tuned on synthetic and human-generated data, analyzing the effects of different prompting methods, and evaluating the performance of multilingual models trained on synthetic datasets.

5.1 N-Shot Prompting

The performance results of GPT-4 for idiomaticity detection across multiple languages in different settings—zero-shot, few-shot with synthetic data, and few-shot with human-generated data—are presented in Table 3. GPT-4 performs the weakest in the zero-shot setting across all languages, with English showing the lowest performance (55.72%). However, performance improves significantly in the few-shot setting, with all languages exhibiting a notable increase in F1 scores. The use of human-generated data yields the highest performance for all languages, and the improvements are consistent across them. These results suggest that GPT-4 benefits significantly from few-shot prompting. Additionally, while human-generated examples lead to the best performance, results with synthetically generated examples are not far behind, in-

[4] In this study, an instance of a PIE that the judge could not classify based on the context were labeled as "unknown."

		EN		JP		IT		TR	
		Train	**Macro Avg. F1**	**Train**	**Macro Avg. F1**	**Train**	**Macro Avg. F1**	**Train**	**Macro Avg. F1**
GPT-4	**Zero-shot**	-	55.72	-	75.36	-	71.80	-	66.39
	Few-shot (w/ synthetic)	-	78.99	-	81.42	-	85.12	-	82.66
	Few-shot (w/ human-generated)	-	83.24	-	83.21	-	87.65	-	86.62
Task-specific models	**mBERT**	GPT-4	75.52±0.5	GPT-4	75.35±0.2	GPT-4	71.71±1.2	GPT-4	70.37±0.4
		VNC-Tokens	84.92±1.7	Open MWE	93.10±0.2	Dodiom	85.36±0.2	Dodiom	82.43±0.6
	XLM-Roberta	GPT-4	77.52±0.8	GPT-4	78.46±1.3	GPT-4	76.44±0.9	GPT-4	76.41±0.5
		VNC-Tokens	84.86±0.5	Open MWE	94.24±0.1	Dodiom	86.35±0.2	Dodiom	85.52±0.2
	DistilBERT	GPT-4	77.27±0.8	GPT-4	57.37±0.8	GPT-4	64.45±0.2	GPT-4	61.75±0.2
		VNC-Tokens	**89.02**±0.1	Open MWE	85.05±0.1	Dodiom	79.20±0.8	Dodiom	74.68±0.3
	Language-specific BERT	GPT-4	77.06±0.6	GPT-4	80.76±0.3	GPT-4	76.56±2.3	GPT-4	78.15±0.6
		VNC-Tokens	88.66±0.8	Open MWE	**94.36**±0.1	Dodiom	**89.22**±0.2	Dodiom	**88.81**±0.5

Table 3: A performance comparison of models trained on synthetically generated datasets, human-generated datasets, and GPT-4, tested using human-generated datasets, with standard errors also provided for task-specific models.

dicating that synthetic data can still be valuable for idiomaticity detection. Furthermore, the number of examples and the order of presentation (figurative-first vs. literal-first) also influence performance (Table 6).

5.2 Task-Specific Fine-tuning

The results of comparing models trained on synthetic data with those trained on human-generated data are presented in Table 3. While task-specific models trained with human-generated data outperform those trained with synthetic data consistently, overall, best results are obtained with DistilBERT in English (89.02%) and language-specific BERTs in Japanese (94.36%), Italian (89.22%) and Turkish (88.81%). Notably, the language-specific BERT model for English also achieve the near-best performances.

The average performance differences based on data source (synthetic vs. human-generated), favoring models trained on human-generated data, are 10 percentage points (pp) for English (ranging from 7 to 12), 19 pp for Japanese (ranging from 14 to 28), 13 pp for Italian (ranging from 10 to 15), and 11 pp for Turkish (ranging from 10 to 13). Additionally, the performance gap between the best-performing synthetically trained model and GPT-4 in zero-shot setting is 22 pp for English (55.72% vs. 77.52% with XLM-Roberta), 5 pp for Japanese (75.36% vs. 80.76% with Japanese BERT), 5 pp for Italian (71.80% vs. 76.56% with Italian BERT), and 12 pp for Turkish (66.39% vs. 78.15% with BERTurk). In each language, the top-performing task-specific models trained on synthetic data is outperformed by GPT-4 in few-shot setting with synthetic data.

The results indicate that, while synthetic data is less effective than human-generated data in helping models distinguish between idiomatic and literal meanings, training smaller task-specific models with synthetic data generated from LLMs is a more efficient approach compared to directly using the LLMs in zero-shot setting. Additionally, using synthetic data is more cost and time efficient than relying on human-generated data. For instance, we generate sentences at a cost of $0.004 each, while Haagsma et al. (2020) reports a cost of $0.04 per sentence using a crowdsourcing approach, making our method 10 times cheaper. Moreover, annotating 100,000 examples in Hashimoto and Kawahara (2009) takes 230 hours for two people, whereas using an LLM can achieve the same task in approximately 45 hours, providing a solution that is 5 times faster. This highlights the significant time and resource savings offered by synthetic data generation, especially given the lengthy and expensive process of human annotation.

5.3 Zero-Shot vs. Enhanced Prompting

To analyze the effect of different prompts used in dataset generation on data quality, the performances of models fine-tuned with samples generated by two distinct prompt types (i.e., zero-shot and enhanced prompting) is analyzed and presented in Table 4. Additionally, the averages

	EN		JP		IT		TR	
	Method	**Macro Avg. F1**	**Method**	**Macro Avg. F1**	**Method**	**Macro Avg. F1**	**Method**	**Macro Avg. F1**
mBERT	Zero-shot	74.31	Zero-shot	74.86	Zero-shot	65.68	Zero-shot	68.78
	Zero-shot filtered	75.04	Zero-shot filtered	74.12	Zero-shot filtered	68.92	Zero-shot filtered	69.48
	Enhanced	75.05	Enhanced	70.07	Enhanced	67.08	Enhanced	67.49
XLM-Roberta	Zero-shot	77.41	Zero-shot	80.46	Zero-shot	71.01	Zero-shot	75.14
	Zero-shot filtered	78.29	Zero-shot filtered	78.50	Zero-shot filtered	69.47	Zero-shot filtered	73.04
	Enhanced	77.97	Enhanced	77.69	Enhanced	72.02	Enhanced	73.62
DistilBERT	Zero-shot	77.29	Zero-shot	59.37	Zero-shot	62.30	Zero-shot	58.73
	Zero-shot filtered	76.57	Zero-shot filtered	57.58	Zero-shot filtered	61.12	Zero-shot filtered	59.67
	Enhanced	**79.19**	Enhanced	54.27	Enhanced	60.81	Enhanced	57.04
Language-specific BERT	Zero-shot	76.70	Zero-shot	80.69	Zero-shot	**77.31**	Zero-shot	77.16
	Zero-shot filtered	76.43	Zero-shot filtered	**80.95**	Zero-shot filtered	73.47	Zero-shot filtered	**77.77**
	Enhanced	76.90	Enhanced	76.86	Enhanced	73.71	Enhanced	77.54

Table 4: A performance comparison of models trained separately using data generated from different prompting settings and evaluated with human-generated datasets.

from three experiments, conducted on randomly selected subsets of 40 samples (referred to as zero-shot filtered) drawn from a total of 60, are provided. The enhanced prompt shows benefits only for English, achieving the highest score of 79.19% by fine-tuning DistilBERT with data from enhanced prompting. However, overall performance differences between the two prompt types are minimal. One possible explanation for the limited improvement from the enhanced prompt is the performance gap between GPT-4's capabilities in English and non-English languages (Ahuja et al., 2023). The sample size does not yield consistent results between the zero-shot and zero-shot filtered prompts; it decreases performance in some models while increasing it in others.

The diversity analysis results (Figure 3) indicate that data samples generated with enhanced prompting generally exhibit greater diversity than those generated by humans or with zero-shot prompting, except for English, where human-generated data demonstrates higher diversity. While the results highlight the effectiveness of enhanced prompting in generating more semantically diverse outputs, the observation that models trained with enhanced prompt-generated data are less successful than those trained with human-generated data suggests that idioms are often used with specific sentence structures in real-world scenarios, rather than with varied sentence structures.

	EN	**JP**	**IT**	**TR**
GPT-4 (zero-shot)	55.72	75.36	71.80	66.39
GPT-4 (few-shot w/ synthetic)	78.99	**81.42**	**85.12**	**82.66**
Monolingual best (w/ synthetic)	77.52	80.76	75.56	78.15
mBERT	77.99	77.12	72.36	71.98
XLM-Roberta	75.19	79.21	77.35	77.00
DistilBERT	77.78	61.15	65.28	64.05
Language-specific BERT	**79.31**	78.56	79.55	79.32

Table 5: A performance comparison of multilingual models trained with merged synthetic datasets from different languages. The results reflect macro average F1 scores. First three rows provide GPT-4 tested in zero-shot and few-shot setting, and monolingual best performances, respectively.

5.4 Multilingual Idiomaticity Detection

The multilingual idiomaticity detection experiments yield notable results when comparing various model architectures across English, Japanese, Italian, and Turkish (see Table 5). In particular, smaller multilingual task-specific models consistently outperform GPT-4 in the zero-shot setting. However, GPT-4 generally performs better when synthetic data is also provided during the test phase (i.e., in the few-shot setting). The only exception is in English, where English BERT achieves 79.31% compared to GPT-4's 78.99% in the few-shot setting. Comparing monolingual and multilingual task-specific models reveals that the best multilingual model generally outperforms the best monolin-

gual model, except for Japanese. This performance disparity suggests that model size alone does not dictate effectiveness in the idiomaticity detection task. Instead, specialized architectures, even if smaller, or different prompting settings can better capture the necessary patterns for identifying idiomatic expressions across various languages.

6 Conclusion

In this article, we create synthetic idiom corpora in multiple languages using GPT-4 and evaluate the effectiveness of models trained on these corpora for the idiomaticity detection task. Additionally, we have analysed the impact of the prompts used during the example generation process on corpus quality and assessed the influence of synthetic data on the performance of multilingual models.

The results indicate that while synthetic data may not match the quality of human-generated data, it offers significant advantages in terms of cost and time efficiency. Furthermore, smaller task-specific models trained on the synthetic data generated by the LLM outperform the LLM itself on the same task in the zero-shot setting. However, the LLM surpasses these models when synthetic data is also provided during the test phase (i.e., in few-shot prompting setting), highlighting the potential of synthetic data to enhance LLM performance. In this setup, the LLM achieves results comparable to using human-generated data in few-shot prompting setting, with a difference of only 2 percentage points for Japanese and Italian and 4 percentage points for English and Turkish.

Our findings also reveal that more complex prompts during the synthetic data generation process do not consistently produce higher-quality examples. These complex prompts produce beneficial results only in English, likely because GPT-4 performs more effectively in English than in other languages across various tasks. Overall, while the LLM's performance in idiomaticity detection remains lower than that of task-specific models, as is the case in other natural language processing tasks, its generalization potential makes it a highly valuable resource.

Future work could focus on more sophisticated prompting methods, refining the reasoning process of the utilized LLM, and expanding the study to include additional languages and LLMs. Additionally, the generated data could be used for instruction-tuning of LLMs to explore potential improvements in their ability to handle idiomatic expressions across diverse languages. Overall, our findings highlight the pivotal role LLMs can play in generating idiom corpora as a cost and time-effective alternative to methods relying on human effort, as well as the effect of synthetic data in enhancing LLM performance on idiomaticity detection task.

Limitations

One notable limitation of the article is that synthetic data generation relied solely on GPT-4. Additionally, we constructed synthetic idiom corpora for English, Italian, Japanese, and Turkish, which limits our scope as these languages might not encompass the full spectrum of idiomatic usage found across all languages. Moreover, our data generation employed two distinct prompting techniques. While these prompts showed promise, further refinement in reasoning process of GPT-4 or exploration of more advanced prompt engineering could enhance data quality. Another consideration is that, given GPT-4's extensive training on a large and diverse corpus, there is potential for data leakage, where the model may have encountered datasets used. Such exposure could affect the diversity and authenticity of the generated samples. Furthermore, our evaluation setup exclusively utilized BERT variants for training task-specific models. However, fine-tuning or instruction-tuning a broader set of models could yield additional insights and highlight model-specific strengths.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. 2022. Vector representations of idioms in conversational systems. *Preprint*, arXiv:2205.03666.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Cristina Cacciari and Patrizia Tabossi. 2014. *Idioms*. Psychology Press.

Anna B. Cieślicka. 2015. Idiom acquisition and processing by second/foreign language learners. In *Bilingual Figurative Language Processing*, pages 208–244. Cambridge University Press.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. *Towards a Shared Task for Multiword Expressions (MWE 2008)*, page 19.

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.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. *Preprint*, arXiv:2205.12679.

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.

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

Yuzhao Heng, Chunyuan Deng, Yitong Li, Yue Yu, Yinghao Li, Rongzhi Zhang, and Chao Zhang. 2024. Proggen: Generating named entity recognition datasets step-by-step with self-reflexive large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, page 15992–16030. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. 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.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, page 11065–11082. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 462–477. Curran Associates, Inc.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam

Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, page 87–94, Berlin, Heidelberg. Springer-Verlag.

Stefan Schweter. 2020a. BERTurk - BERT models for Turkish.

Stefan Schweter. 2020b. Italian BERT and ELECTRA models.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. 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).

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *Preprint*, arXiv:2303.04360.

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*, page 2715–2726. Association for Computational Linguistics.

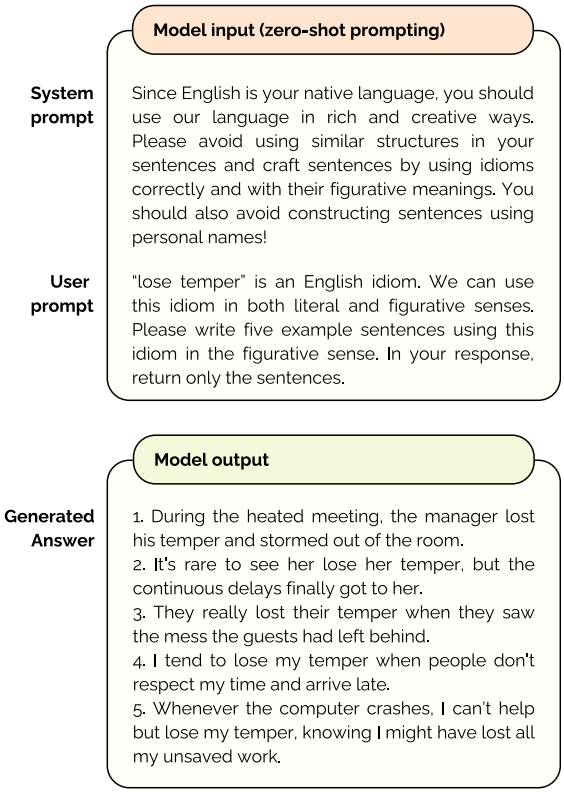

Figure 1: Illustration of a zero-shot prompting setup. The model input consists of a system prompt guiding idiomatic usage, followed by a user prompt requesting example sentences using an idiom in a figurative sense. The model output provides five example sentences following these instructions.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *Preprint*, arXiv:2109.09193.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

A Additional Figures

This appendix provides supplementary materials to enhance the understanding of the experimental setups and results presented in the main text. The figures and table included here illustrate key aspects of the data generation methods, evaluation settings, and diversity analysis.

Data Type	Setting	Same PIE?	# Examples	Order of Examples	Macro Avg. F1
Synthetic	Zero-shot	-	0	-	55.72
	One-shot	No	1	Figurative-first	51.16
		Yes	1	Figurative-first	65.57
	Few-shot	Yes	3	Figurative-first	74.96
		Yes	3	Literal-first	78.99
		Yes	5	Figurative-first	73.90
Human-generated	Few-shot	Yes	3	Literal-first	**83.24**

Table 6: Analysis of the effect of using synthetic or human-generated data, the number of examples, the order of examples, and whether the examples contain the same PIE as the test sentence in evaluating GPT-4 for English.

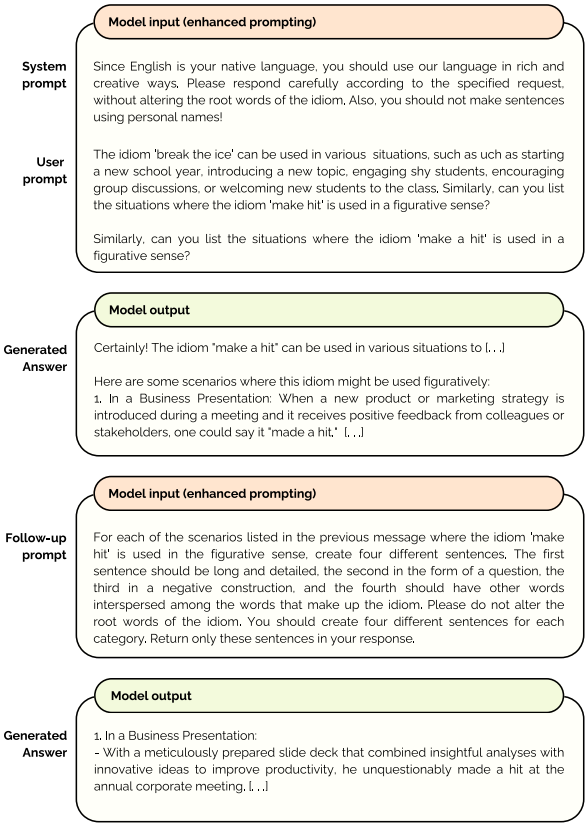

Figure 2: Illustration of an enhanced prompting setup. The model is instructed to explore an idiom in various scenarios, while following specific linguistic constraints. For each scenario, the model generates four unique sentences. Ellipsis ([...]) indicates omitted sections for brevity.

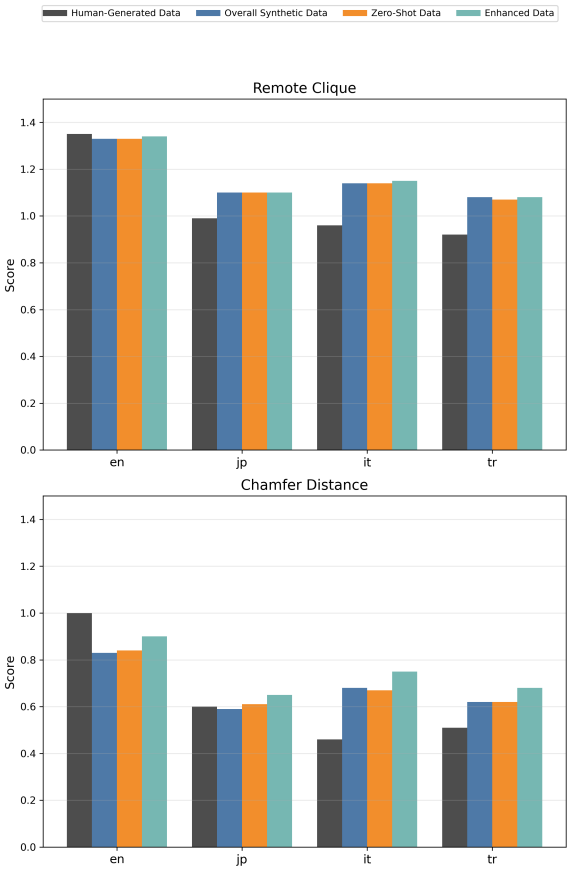

Figure 3: Comparison of the diversity between human-generated data and synthetic data produced using zero-shot and enhanced prompting, evaluated using remote clique score and Chamfer distance score. For both metrics, higher scores indicate greater diversity.