A Unified Turkic Idiom Understanding Benchmark: Idiom Detection and Semantic Retrieval Across Five Turkic Languages

Gözde Aslantaş[1,2], Tunga Güngör[1]

[1]Boğaziçi University, Department of Computer Engineering, Istanbul, Türkiye
gozde.aslantas@std.bogazici.edu.tr,
gungort@bogazici.edu.tr
[2]Yapı Kredi Teknoloji A.Ş., Applied AI and R&D, Istanbul, Türkiye
gozde.aslantas@ykteknoloji.com.tr

Abstract

Idiomatic expressions are culturally grounded, semantically opaque, and challenging for multilingual natural language processing systems. Despite the large speaker population of Turkic languages, resources for monolingual and cross-lingual idiom understanding remain scarce. We introduce the first unified benchmark for idiom understanding across Turkish, Azerbaijani, Turkmen, Gagauz, and Uzbek, featuring token-level idiom span annotations. We evaluate seven models for idiom identification and nine embedding models for semantic retrieval under multiple fine-tuning schemes. Our benchmark enables systematic analysis of how idiomatic meanings are shared, transformed, or diverge across Turkic languages.

1 Introduction

Idiomatic expressions are semantically complex, non-compositional, culturally grounded, and often language-specific, making them difficult for multilingual models to interpret. Robust idiom understanding requires cultural knowledge and context-sensitive reasoning, and remains a challenge despite the advances in multilingual natural language processing (NLP).

Turkic languages, spoken by over 170 million people (Johanson and Csató, 2015; Eberhard et al., 2025), are notably underexplored in this area. Prior work has focused on Turkish or isolated bilingual lists, but lacks a unified resource that covers multiple Turkic languages and that supports idiom span detection, cross-lingual semantic retrieval, systematic evaluation of pretrained and fine-tuned models, and controlled LLM-based idiom reasoning experiments.

To address this gap, we introduce the first comprehensive **Turkic Idiom Understanding Benchmark**, covering the Turkish (TR), Azerbaijani (AZ), Turkmen (TK), Gagauz (GA), and Uzbek (UZ) languages. The benchmark supports the following tasks:

1. **Idiom Span Detection**: A BIO-tagged token classification task where the models identify idiomatic multi-word expressions in context, enabling evaluation of idiom span recognition across languages.

2. **Semantic Retrieval**: A multilingual idiom-meaning retrieval suite with three tasks: cross-lingual idiom-to-idiom retrieval, monolingual idiom-to-meaning retrieval, and cross-lingual idiom-to-meaning retrieval.

We use seven models for idiom span detection and nine models for semantic retrieval under different fine-tuned schemes. We also augment the retrieval analyses with **LLM-based semantic validation** to assess the figurative alignment and the cross-lingual consistency.

Our contributions are as follows:

(1) A unified Turkic idiom dataset. We release the first large-scale resource covering five Turkic languages, featuring aligned idioms, meanings, example sentences, and BIO-tagged span annotations.

(2) A dual-pipeline benchmark. We jointly evaluate idiomatic span detection and semantic retrieval, enabling an analysis of how surface-level idiomatic identification impacts semantic understanding.

(3) Fine-tuning strategies. We systematically compare pretrained encoders in base, TR-only fine-tuned, multilingual fine-tuned, and language-specific fine-tuned settings across three retrieval tasks.

(4) LLM idiom evaluation. We introduce controlled prompts for idiom and meaning validation.

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 38–51
March 29, 2026 ©2026 Association for Computational Linguistics

(5) A full reproducible benchmark. We release all the code, datasets, and evaluation scripts for reproducibility.[1]

2 Related Work

Idiom Understanding in Language Models. Large transformer encoders capture layered syntactic and semantic signals (Clark et al., 2019; Wu et al., 2020; Ethayarajh, 2019), supporting context-based idiom interpretation. Yet, multilingual idiom competence remains an open issue. Tedeschi et al. (2022) show that multilingual BERT detects idiomatic usage, but fails on literal senses under insufficient contextual diversity. Tayyar Madabushi et al. (2021) report that monolingual models capture syntactic features, but multilingual encoders, such as XLM-R, demonstrate only limited cross-lingual generalization. These findings highlight that architecture alone is insufficient and robust idiom modeling requires diverse and semantically grounded pre-training.

Recent work explores leveraging LLMs to accelerate the construction of idiom corpora. Arslan et al. (2025) propose generating synthetic idiom corpora with LLMs to reduce expert annotation cost and time. While models trained on synthetic data underperform compared to human-curated corpora, the synthetic approach offers significant cost efficiency and improves LLM performance under few-shot prompting. This suggests promising avenues for scalable creation of idiom datasets, complementing traditional corpus-building efforts.

Idioms in Turkic Languages. Research on Turkic idioms is predominantly linguistic rather than computational, focusing on cultural, semantic, and comparative descriptions for Turkish (Güven, 2020; Karakus, 2020; Püsküllüoğlu, 2012) and across Turkic branches (Yeşildere Aldan, 2020; Bayramov, 2014; Goçgeldiyeva, 2018). Uzbek and Gagauz resources remain limited to descriptive lexicography (Rahmatullayev, 1978; Abdullayeva, 2015; Mihail, 2020), with no machine-readable corpora or NLP benchmarks. Overall, the field lacks annotated idiom datasets and cross-lingual evaluation suites for the Turkic family.

More recent computational efforts have begun addressing these gaps. Umut et al. (2025) systematically evaluate encoder-only and decoder-only LLMs for idiomaticity detection and idiom identification in Turkish, showing that fine-tuned encoder-based models (notably mDeBERTa-V3) outperform decoder-based models in supervised settings, though few-shot prompting allows models like OpenAI-o3 to approach similar performance levels. This highlights both the advantage of supervised fine-tuning and the emerging potential of prompting-based adaptation for idiom-related tasks in low-resource languages.

Beyond idioms, other forms of figurative language have been computationally modeled in Turkic languages. Inan (2025) introduce a contrastive retrieval framework for metaphor detection in Turkish, achieving high recall using dense and contrastive semantic models. Similarly, Biyik et al. (2024) present the first Turkish euphemism dataset, expanding figurative language research to cover euphemistic expressions and benchmarking transformer-based models on euphemism detection.

In Uzbek, Abjalova et al. (2025) explore stylistic identification of idiomatic units through an NLP-driven approach using morphological analysis and contextual embeddings (SBERT-uz). Their work enables automatic classification of idioms by speech style (e.g., formal, colloquial), representing one of the earliest attempts to computationally model idiomatic stylistics in Uzbek.

Collectively, these studies highlight a significant transition from descriptive linguistics to the computational processing of idioms and figurative language within the Turkic family, emphasizing the importance of dataset creation and model adaptation for resource-scarce languages.

3 Dataset Construction

We construct a unified cross-lingual idiom lexicon for Turkish, Azerbaijani, Turkmen, and Gagauz, and later extend it to include Uzbek, supporting idiom understanding and retrieval across under-resourced Turkic languages. The pipeline involves: (i) adopting a manually aligned multilingual idiom inventory for four closely related Turkic languages (TR–AZ–TK–GA) from prior linguistic work, (ii) expanding the lexicon by collecting and normalizing Uzbek idioms from diverse lexicographic sources, (iii) aligning Uzbek idioms to existing entries via semantic similarity modeling using multilingual sentence embeddings. In this way, idioms that share equivalent meanings across Turkic varieties are aligned while preserving their language-

[1] https://github.com/gozdeaslantas/Turkic_Idiom_Understanding_Benchmark

specific surface forms and cultural nuances.

3.1 Turkic Idiom Lexicon (TR-AZ-TK-GA)

We begin by constructing a high-quality idiom lexicon covering four Turkic languages: Turkish (TR), Azerbaijani (AZ), Turkmen (TK), and Gagauz (GA). As the primary source, we use the work of Yeşildere Aldan (2020), which provides a manually aligned inventory of idiomatic expressions shared across these varieties.

While this inventory offers aligned idioms and partial glosses, it lacks consistent definitions and contextual examples. To create a lexicon including this information, we use a controlled LLM-based normalization step (GPT-4.1) to generate standardized forms, monolingual definitions, and example sentences for each idiom within its own language; not across languages. Thus, the LLM is not responsible for performing alignment, but rather for within-language enrichment of already aligned idioms. To ensure reliability, we apply an automated quality control step in which a separate LLM-based verifier (GPT-5.1) evaluates each generated entry along three criteria: semantic fidelity to the idiom, internal consistency between definition and example, and idiomatic (non-literal) usage in context. Entries that fail any criterion, exhibit contradictions, or receive low confidence scores are discarded through rule-based filtering. The resulting lexicon thus contains only validated idiom entries with consistent definitions and examples. In the preliminary work, we tested a number of prompts, and we empirically selected the one that consistently produced coherent and semantically faithful definitions across all languages. The prompts can be found in Appendix A.

3.2 Uzbek Idiom Collection

To extend the resource beyond the four languages, we compile a set of Uzbek idioms from multiple authoritative and heterogeneous sources. Our primary reference is the monolingual idiom dictionary of Rahmatullayev (1978). To broaden coverage and capture contemporary and web-based usage, we additionally extract idiomatic expressions from the publicly available Toshquvvatov Idiom Dictionary[2] and from idiom lists provided in the UzSchoolCorpora platform[3].

Language	#Idioms	with Def&Exp	Idiom Annot.
Turkish (TR)	3456	3446	843
Azerbaijani (AZ)	3206	3204	685
Turkmen (TK)	1134	1087	232
Gagauz (GA)	140	60	83
Uzbek (UZ)	280	280	51

Table 1: Dataset statistics per language: numbers of idioms, idioms with definition and example, and idioms with an annotated sentence.

3.3 Unified Cross-Lingual Representation

To enable cross-lingual alignment between the idioms in the four languages and the Uzbek idioms, we construct language-independent sense representations. For each idiom in the first set, we concatenate its monolingual glosses (TR, AZ, TK, GA) into a string. For Uzbek idioms, we use the monolingual Uzbek definition and, when available, an English gloss. These representations are subsequently embedded into a shared multilingual sentence embedding space. The alignment is obtained using cosine similarity between the embeddings.

The statistics about the compiled lexicon are shown in Table 1. The lexicon contains a total of 3,469 idioms with distinct definitions in JSONL format. We note that Turkish and Azerbaijani have a corresponding idiom in majority of these idioms. On the other hand, the other three languages, especially Gagauz and Uzbek have a corresponding idiom in less than 10% of the idioms.

4 Methodology

4.1 Idiom Span Detection

In the idiom span detection task, we evaluate how effectively pretrained transformers recognize idiomatic multi-word expressions in a sentence in Turkic languages. Each model is provided with the example sentence of the idiom as input and is required to detect the token span corresponding to the idiomatic expression appearing in that sentence. We compare multilingual (XLM-R, mBERT), Turkish-specific (BERTurk, ELECTRA-TR, DistilBERT-TR, ConvBERT-TR), and embedding-oriented (E5-Large) encoder models. In this task, we address the following research questions:

- **RQ1:** Can Turkish-trained idiom span detectors generalize to related Turkic languages in a zero-shot setting?

- **RQ2:** How many examples are required to adapt to each Turkic language? Does few-

[2] https://toshquvvatovdictionary.wordpress.com/ozbekcha-iboralar/

[3] https://uzschoolcorpara.uz/

shot learning saturate quickly or require more supervision?

- **RQ3:** Which model architectures encode idiomatic structure most effectively?

Together, the experiments conducted in this task form the first large-scale idiom span evaluation benchmark across five Turkic languages.

4.1.1 Span-Level Annotation

Idioms in Turkic languages often exhibit multiword, morphologically rich, and occasionally noncontiguous structures. To train idiom span detection models, we first transformed all idiom occurrences into token-level BIO labels.

All idioms with examples were annotated with the `B-IDIOM` label for the first sub-token and `I-IDIOM` for all subsequent sub-tokens. Idiom spans were located in sentences using fuzzy string matching at the word level. The method applied normalization and stemming to both idiom and sentence tokens, followed by a hierarchical matching procedure involving exact token equality, approximate matching based on a Levenshtein edit distance of at most one, and substring inclusion checks to capture minor inflectional or orthographic variations. Instances that could not be reliably aligned under these criteria were eliminated. Matched word-level spans were then projected onto subword tokens, and BIO labels were propagated to all corresponding sub-tokens, ensuring consistent token-level supervision across tokenization schemes. Table 1 (last column) reports the number of examples annotated with BIO labels.

4.1.2 Idiom Span Detection Pipeline

We first train all models as token-level sequence taggers on the annotated Turkish dataset. Each model is then evaluated separately on four Turkic languages: Azerbaijani, Turkmen, Gagauz, and Uzbek. This setup enables us to measure structural and idiomatic similarity across the family by observing to which languages the Turkish-trained model generalizes without additional supervision.

Following the zero-shot evaluation, we explore cross-lingual transfer by fine-tuning the Turkish-trained models on small supervised subsets (10, 30, 50 examples, and the full training set) for each target language (AZ, TK, GA, UZ), then evaluating on the full test split of that language. This setup quantifies data requirements for idiom span recognition, highlights which architectures adapt effectively with minimal supervision, and examines how "close" each language is to Turkish in the idiomatic sense.

The overall experimental design thus allows for a fine-grained linguistic analysis of idiom similarity across Turkic languages in the following senses:

1. **Zero-shot evaluation** (TR model for AZ/TK/-GA/UZ) reveals inherent structural closeness encoded by the pretrained models.

2. **Few-shot adaptation** (10/30/50/full) reveals how quickly each language adapts, indicating the underlying typological distance from Turkish.

3. **Model comparison** shows which architectures (multilingual vs. Turkish-only vs. embedding-based) induce better cross-lingual inductive bias.

This multi-stage pipeline offers a novel approach to measuring idiom-level linguistic proximity within the Turkic language family.

4.2 Idiom Semantic Retrieval

In the idiom semantic retrieval task, we evaluate whether the encoder models can align the idioms and their meanings in monolingual and cross-lingual settings. We address the following research questions:

- **RQ1:** To what extent can supervision from Turkish idiom-meaning pairs transfer to other Turkic languages in zero-shot settings?

- **RQ2:** Does monolingual fine-tuning on idiom-meaning data improve retrieval performance in that language?

- **RQ3:** Does multilingual supervision across all languages yield better idiomatic alignment than single-language fine-tuning?

In the tasks explained in Sections 4.2.1 through 4.2.3, for each encoder model, we use the pretrained model and three fine-tuning strategies shown below. All fine-tuned checkpoints can be found in the repository.

TR-Only Fine-Tuning fine-tunes the model using Turkish idiom-meaning pairs.

Monolingual Fine-Tuning trains independent models for each language using its own idiom-meaning pairs, enabling evaluation of language-specific adaptation.

Cross-lingual Fine-Tuning trains on idiom-meaning pairs where the idiom is from Turkish and the meaning is from one of the other four languages, repeated for each of the four languages. This setup encourages a shared semantic space that supports retrieval of meanings in one language using idioms from another.

We implement a unified `SemanticEncoder` wrapper that supports both SentenceTransformer models and HuggingFace encoders. For each idiom/meaning, we encode it using the model and return normalized embeddings. We use five pooling strategies to obtain a single embedding corresponding to the idiom/meaning: CLS embedding (last layer), mean pooling of the tokens (last layer), max pooling of the tokens (last layer), mean pooling of the tokens (sum of first and last layers), and mean pooling of the tokens (weighted sum of last four layers). Cosine similarity is used to match the idiom/meaning embeddings. Each model-pooling combination is tested independently across all tasks and the score of the best pooling strategy is reported per model.

4.2.1 Task 1: Cross-Lingual Idiom Retrieval

This task evaluates whether embedding models can align idiomatic expressions across closely related Turkic languages. Given a Turkish idiom, the goal is to retrieve its semantically equivalent idiom from a candidate set in Azerbaijani, Turkmen, Gagauz, and Uzbek. Retrieval is performed by encoding the input idiom and all candidates using an encoder and ranking them by cosine similarity.

We compare three configurations: (a) pretrained model, (b) TR-only fine-tuned model, and (c) cross-lingual fine-tuned model for each language. The task analyzes whether Turkish supervision alone is sufficient to induce cross-lingual alignment and multilingual training provides additional benefits.

4.2.2 Task 2: Monolingual Idiom-to-Meaning Retrieval

This task assesses whether an embedding model can retrieve the correct dictionary-style meaning of an idiom within the same language. In contrast to the cross-lingual settings, this task isolates within-language semantic resolution, probing whether the model can infer figurative meaning from idiomatic form.

For each language (TR, AZ, TK, GA, UZ), we treat an idiom as input and rank all possible meanings in that language as candidates. Success requires the correct meaning to be ranked highest. We report standard retrieval metrics.

We compare two configurations: (a) pretrained model and (b) monolingual fine-tuned model. This setup enables us to measure when monolingual supervision strengthens semantic representations, and when it fails or harms the performance.

4.2.3 Task 3: Cross-Lingual Idiom-to-Meaning Retrieval

This task evaluates whether an embedding model can retrieve the correct meaning of a Turkish idiom when the candidate meanings are drawn from a different Turkic language (AZ, TK, GA, UZ). Unlike Task 1 (idiom-to-idiom), we assess how well the models capture cross-lingual semantic interpretation, not just surface-level lexical alignment.

We compare three configurations: a) pretrained model, b) TR-only fine-tuned model, and c) cross-lingual fine-tuned model for each language. Given a Turkish idiom, the goal is to identify the correct meaning among all definitions in the other language. Standard retrieval metrics are used for evaluation.

This task investigates whether idiomatic semantics learned in Turkish transfer effectively across related languages and whether multilingual supervision yields gains beyond single-language fine-tuning.

4.3 LLM-Based Semantic Evaluation

To validate the reliability of our gold annotations in the lexicon and to complement the retrieval-based experiments, we use an GPT-4.1 to judge whether a given idiom-meaning pair reflects the intended figurative interpretation. The model assigns a quality label (*HIGH / MEDIUM / LOW*) and provides a brief justification. Prompt templates and output format are given in Appendix B.

Monolingual Evaluation In the monolingual setting, the LLM assesses whether the meaning correctly captures the idiom within the same language. This relates to Task 2 by verifying that evaluation labels correspond to real idiomatic semantics rather than literal paraphrasing or annotation noise.

Cross-Lingual Evaluation In the cross-lingual setting, the GPT-4.1 evaluates whether the meaning in another Turkic language preserves the idiom's figurative sense. This relates to Task 1 and Task 3, providing evidence on how cultural grounding and semantic transfer impact cross-lingual idiom interpretation.

5 Experiments and Results

We conduct two sets of experiments: (i) idiom span detection for evaluating cross-lingual sequence tagging in the Turkic family, and (ii) idiom semantic retrieval for probing cross-lingual idiom and meaning alignment. This section describes the experimental setup, baselines, ablation studies, and both quantitative and qualitative results.

5.1 Idiom Span Detection

In this section, we quantify how well token-level models trained on Turkish idiom spans transfer to other Turkic languages: Azerbaijani (AZ), Turkmen (TK), Gagauz (GA), and Uzbek (UZ). We focus on span-level detection of idioms, modeled as a BIO token classification task.

All models are fine-tuned on Turkish idiom span annotations and then (1) evaluated zero-shot on AZ/TK/GA/UZ, and (2) adapted with few-shot supervision (10, 30, 50, full training examples per language). We consider both multilingual and Turkish-based encoders: XLM-R, mBERT, BERTurk, ELECTRA-tr, DistilBERT-tr, ConvBERT-tr, and Turkish-E5-Large.[4]

5.1.1 Experimental Setup

Training Configuration All models are trained using the same configuration. The details are provided in Appendix C. Each experiment is repeated with three different random seeds, and results are reported as averages over these runs.

Data Splits For idiom span detection, we use the idioms with gold span annotations reported in the last column of Table 1. Data splits are created by randomly shuffling idioms while ensuring that all sentence instances corresponding to the same idiom remain in the same split. We allocate 15% of the idioms to the test set and split the remaining data into 90% for training and 10% for validation. The validation set is used for hyperparameter tuning and model selection (e.g., determining optimal learning rate, batch size, and early stopping based on validation loss), while the final reported results are obtained from the test set.

Model	P	R	F1	Acc
XLM-R	0.751	0.766	0.758	0.913
mBERT	0.701	0.751	0.725	0.902
BERTurk	0.827	0.817	0.822	0.938
ELECTRA-tr	0.876	0.878	0.877	0.953
DistilBERT-tr	0.622	0.664	0.642	0.885
ConvBERT-tr	0.884	0.875	0.880	0.951
Turkish-e5-L	0.832	0.842	0.837	0.947

Table 2: Idiom detection results on TR test set with all models fine-tuned on TR only. (Turkish-e5-L: Turkish-E5-Large)

Zero-Shot Evaluation We treat TR as the sole training language and evaluate the resulting models directly on test sets from the other four languages without any target supervision. This setup assesses how idiom span knowledge transfers across the Turkic family.

Few-Shot Adaptation We further fine-tune each TR-fine-tuned checkpoint on the target language using $n \in \{10, 30, 50\}$ labeled training instances, as well as the full training set of the target language. We report the best F1 achieved among these settings for each model/language pair.

Evaluation Metrics We report token-level precision, recall, F1, and accuracy. F1 is the primary model selection metric and is used for all comparisons. Cross-lingual aggregated results are presented both per language (AZ/TK/GA/UZ) and macro-averaged.

5.1.2 Monolingual TR Performance

Before the zero-shot and few-shot evaluations, we test the TR-fine-tuned models on Turkish idioms to form an upper performance limit for cross-lingual experiments. Table 2 reports token-level results on the TR test set after fine-tuning for five epochs. All models reach a strong idiom span F1 on Turkish (0.64-0.88), confirming that the BIO formulation is learnable with our annotation scheme.

Across the Turkish-centric encoders, **ConvBERT-tr** and **ELECTRA-tr** obtain the strongest TR performance (0.880 and 0.877 F1), surpassing the more standard BERTurk and DistilBERT-tr baselines. While multilingual models such as XLM-R and mBERT lag behind (0.758 and 0.725 F1), **Turkish-E5-L**

[4] xlm-roberta-base, bert-base-multilingual-cased, dbmdz/bert-base-turkish-cased, dbmdz/electra-base-turkish-cased-discriminator, dbmdz/distilbert-base-turkish-cased, dbmdz/convbert-base-turkish-mc4-cased, ytu-ce-cosmos/turkish-e5-large.

Model	AZ	TK	GA	UZ	Avg
XLM-R	0.69	0.26	0.67	0.61	0.56
mBERT	0.63	0.36	0.45	0.35	0.45
BERTurk	0.35	0.23	0.89	0.04	0.38
ELECTRA-tr	0.33	0.29	0.97	0.26	0.46
DistilBERT-tr	0.22	0.18	0.49	0.08	0.25
ConvBERT-tr	0.48	0.55	0.87	0.08	0.49
Turkish-e5-L	0.76	0.42	0.90	0.56	0.66

Table 3: Zero-shot idiom detection results (F1) on AZ/TK/GA/UZ with all models fine-tuned on TR only.

Model	AZ	TK	GA	UZ
XLM-R	0.79	0.62	1.00	1.00
mBERT	0.75	0.63	0.84	0.90
BERTurk	0.72	0.65	1.00	0.92
ELECTRA-tr	0.74	0.68	0.97	0.90
DistilBERT-tr	0.55	0.57	0.60	0.90
ConvBERT-tr	0.80	0.76	0.94	0.90
Turkish-e5-L	0.83	0.72	0.94	1.00

Table 4: Few-shot idiom detection results (F1) for each model and language (maximum over 10/30/50/full settings).

performs competitively (0.837 F1), indicating that embedding-oriented pretraining can still capture span level idiomatic boundaries. Overall, these results suggest that **pretraining on Turkish data remains a decisive advantage** for idiom span detection, forming a strong initialization for subsequent cross-lingual transfer.

5.1.3 Zero-Shot Evaluation

We next freeze each TR-fine-tuned model and evaluate it **zero-shot** on AZ, TK, GA, and UZ without any target-language supervision. Table 3 summarizes the results.

Zero-shot transfer from Turkish follows a clear typological gradient. **Generalization to Azerbaijani (TR to AZ) is consistently strongest** across models (e.g., XLM-R: 0.69 and Turkish-e5-L: 0.76), reflecting close linguistic proximity. Performance degrades substantially for **Turkmen and Uzbek**, where F1 scores are markedly lower and more variable (e.g., DistilBERT-tr: 0.18 on TK, 0.08 on UZ), indicating limited zero-shot transfer to more distant Turkic branches.

Among all models, **Turkish-e5-L** (0.66) and **XLM-R** (0.56) attain the highest macro-average, suggesting that multilingual pretraining improves cross-lingual robustness under zero-shot conditions. Results on **Gagauz** often appear high but should be interpreted cautiously due to the small evaluation set.

Overall, these findings suggest that **zero-shot TR-trained models generalize reliably only to the closest related language (AZ)**, while additional supervision is necessary for robust idiom span detection in more distant Turkic languages.

5.1.4 Few-Shot Adaptation

We evaluate whether limited target-language supervision improves idiom span detection by adapting each TR fine-tuned model using $n \in \{10, 30, 50\}$ examples, as well as the full training set. Table 4 reports the best F1 score achieved for each model and language across these settings.

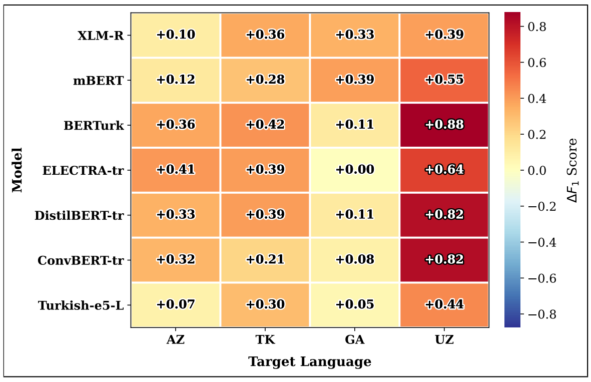

Figure 1: Absolute F1 change (Δ = few-shot F1 − zero-shot F1) for each model and target language (AZ, TK, GA, UZ) in idiom span detection. Warmer colors indicate larger gains from few-shot adaptation.

Few-shot adaptation yields **consistent gains** over zero-shot transfer across all target languages, with the largest improvements observed for **Turkmen and Uzbek**. In these languages, even a small number of in-language examples substantially reduces cross-lingual transfer gaps, while performance typically peaks under full-shot training. **Azerbaijani** shows more moderate gains, consistent with its closer typological proximity to Turkish, whereas results for **Gagauz** should be interpreted cautiously due to the small evaluation set and early saturation effects. Across models, strong zero-shot baselines (e.g., Turkish-e5-L and XLM-R) also benefit from few-shot adaptation, indicating that **limited in-language supervision is sufficient to significantly improve cross-Turkic idiom span detection**. Figure 1 summarizes the improvement from zero-shot to best few-shot performance, with learning curves provided in Appendix D.

5.2 Idiom Semantic Retrieval

We evaluate nine pretrained encoders, including general multilingual models (XLM-R, mBERT), multilingual retrieval-oriented sentence encoders trained with contrastive objectives (MPNet (Reimers and Gurevych, 2019) , Multilingual-E5 (Wang et al., 2024)), Turkish-specific models (BERTurk, ELECTRA-tr, DistilBERT-tr, ConvBERT-tr (Schweter, 2020)), and a Turkish retrieval-oriented sentence encoder (*Turkish-E5-L*) (Izdas et al., 2025), which is a contrastively fine-tuned variant of *multilingual-e5-large-instruct* trained on Turkish retrieval datasets. The *with Def&Exp* column of Table 1 indicates the number of idioms used in the semantic retrieval experiments. Idioms are shuffled and split into training (80%) and test (20%) sets, ensuring all cross-lingual variants with the same meaning remain in the same splits.

For each model, the pretrained configuration and three supervision regimes are employed: (i) TR-only fine-tuned, (ii) monolingual fine-tuned, and (iii) cross-lingual fine-tuned. Models are evaluated under multiple pooling strategies, and results are reported using standard retrieval metrics. Training details are given in Appendix E.

5.2.1 Task 1: Cross-Lingual Idiom Retrieval

Figure 2 shows NDCG performance for retrieving idiom equivalents in AZ, TK, GA, and UZ using a Turkish idiom. The best results are achieved by multilingual retrieval-oriented encoders, with *Multilingual-E5* (0.693) and *Turkish-E5-L* (0.694) outperforming monolingual models BERTurk (0.302) and DistilBERT (0.310), as well as general-purpose multilingual encoders like XLM-R (0.346).

TR-only semantic supervision yields clear gains for several models, notably *Turkish-E5-L* (0.694 to 0.751), *MPNet* (0.479 to 0.628), and *mBERT* (0.512 to 0.579), underscoring the benefit of idiom–meaning alignment. In contrast, *Multilingual-E5* shows no improvement (0.693 to 0.672), suggesting a strong pretrained cross-lingual space.

Pooling matters: mean pooling of the tokens (sum of first and last layers) performs best for the strongest models, suggesting idiomatic meaning is distributed across layers rather than localized in the CLS token. Overall, the results indicate that **retrieval-oriented multilingual encoders form the strongest foundation**, while TR-only supervision offers targeted gains for models not already optimized for semantic retrieval.

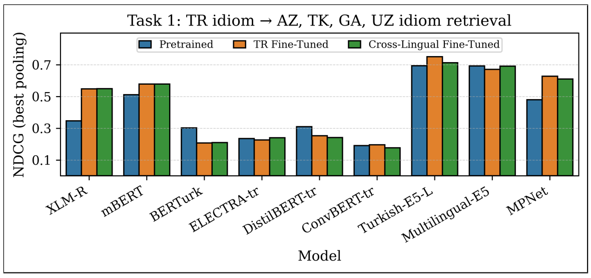

Figure 2: Task 1 results with the best-performing pooling strategy for each model. Scores represent NDCG averaged over AZ/TK/GA/UZ.

Model	Pretrained	Fine-Tuned	Δ
XLM-R	0.16 (M-L)	0.22 (M-FL)	+0.07
mBERT	0.20 (M-L)	0.23 (M-FL)	+0.03
BERTurk	0.25 (M-L)	0.16 (M-L)	–0.09
ELECTRA-tr	0.16 (M-L)	0.17 (M-L)	+0.01
DistilBERT-tr	0.20 (M-L)	0.18 (M-L)	–0.01
ConvBERT-tr	0.15 (M-L)	0.15 (M-L)	+0.00
Turkish-E5-L	0.38 (M-L)	0.60 (M-L)	+0.21
Multi-E5*	0.33 (M-L)	0.46 (M-L)	+0.13
MPNet	0.28 (M-L)	0.41 (M-L)	+0.13

Table 5: Best NDCG scores macro-averaged over TR, AZ, TK, GA, and UZ for Task 2. Pooling abbreviations: **M-L** = mean pooling (last layer), **M-FL** = mean pooling (first+last layers). *Multi-E5 refers to the Multilingual-E5.

5.2.2 Task 2: Monolingual Idiom-to-Meaning Retrieval

Table 5 reports macro-averaged NDCG results averaged over all five languages (TR, AZ, TK, GA, UZ). Fine-tuning yields the largest gains for retrieval-oriented multilingual encoders, with Turkish-E5 (+0.21), followed by MPNet and Multilingual-E5 (+0.13 each). General-purpose multilingual models (XLM-R, mBERT) show only modest gains.

In contrast, monolingual Turkish transformers show minimal change, suggesting that understanding idiomatic meaning requires more than single-language pretraining.

Wilcoxon signed-rank tests show that language-specific fine-tuning significantly improves most models ($p < .05$), with the greatest gains observed in Turkish-E5-L, Multilingual-E5, and MPNet. In contrast, ConvBERT-tr and XLM-R show non-significant effects. Overall, fine-tuning is beneficial but varies in effectiveness, with semantic retrieval architectures consistently outperforming generic multilingual language models. Detailed statistics are in Table 6 in Appendix F.

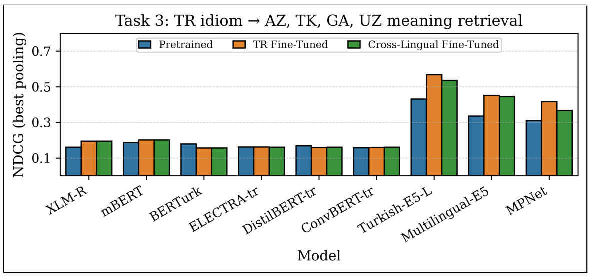

Figure 3: Task 3 results with the best-performing pooling strategy for each model. Scores represent NDCG averaged over AZ/TK/GA/UZ.

5.2.3 Task 3: Cross-Lingual Idiom-to-Meaning Retrieval

Figure 3 shows that cross-lingual idiom-meaning retrieval remains challenging. Pretrained multilingual encoders such as E5-multi and MPNet score only 0.33 and 0.31 NDCG, respectively, while monolingual Turkish models remain around 0.16-0.18. TR-only semantic fine-tuning leads to the most significant improvements, with E5-tr rising from 0.431 to 0.567 (+0.136) and MPNet from 0.309 to 0.416 (+0.107), showing effective generalization across the Turkic language family.**Multilingual fine-tuning offers minimal benefits** (e.g., E5-tr: 0.567 to 0.536, MPNet: 0.416 to 0.367), suggesting diminishing returns once Turkish semantic structure has reshaped the embedding space. TR-only supervision is the key factor in transferring idiomatic meanings across languages.

5.3 LLM Evaluation Results

Evaluation with **GPT-4.1** indicates that the dataset is semantically reliable at the monolingual level. Across all five Turkic languages, **92.8%** of idiom-meaning pairs are rated *HIGH*, with fewer than 4% labeled *LOW*, suggesting limited annotation noise.

With only **70%** *HIGH* ratings, cross-lingual results reflect both linguistic variation and reduced LLM sensitivity, explaining weaker Task 1 and Task 3 performance.

6 Conclusion

We present the first unified benchmark for idiom span detection, semantic retrieval, and cross-lingual alignment across five Turkic languages. Our results show that TR-trained span detectors transfer effectively to closely related languages and achieve competitive performance with limited few-shot supervision. For idiom semantic retrieval, the strongest gains arise from TR-only idiom-meaning supervision, particularly for semantically grounded multilingual encoders such as E5 and MPNet, while monolingual transformers benefit less. Additional multilingual fine-tuning yields only modest improvements. Our findings show that idiomatic meaning in Turkic languages is partly shared and best captured by semantically enriched multilingual representations.

As a future work, we plan to extend the benchmark to additional Turkic languages, such as Kazakh, Kyrgyz, and Uyghur, and complement LLM-based evaluation with human annotation for lower-resource languages.

Ethics Statements

We foresee no ethical concerns related to the methods outlined in this paper.

Limitations

While this work presents the first unified idiom lexicon covering multiple Turkic languages, it has several limitations.

The dataset is imbalanced across languages, particularly with few idiom instances for Gagauz and Uzbek due to limited sources. This may impact the generalizability of evaluation results for these languages.

While idioms in Turkish, Azerbaijani, Turkmen, and Gagauz come from a manually aligned lexicon (Yeşildere Aldan, 2020), the extension to Uzbek uses automatic alignment and data generation with LLMs. Despite automated checks, LLM performance in low-resource languages like Uzbek remains suboptimal. To assess the quality of idioms, we conducted a targeted human evaluation of sampled LLM-aligned idioms for semantic adequacy and usage. This provides an initial quality indicator but cannot replace full expert curation.

Finally, for idiom span detection, we used LLM-generated synthetic sentences, which may not capture the full linguistic diversity of natural idioms. Future research should emphasize broader human validation and the inclusion of expert-annotated corpora, particularly for low-resource Turkic languages

Acknowledgments

We gratefully acknowledge Yeşildere Aldan (2020) for providing the idiom inventory underlying our Turkic idiom lexicon.

References

Gulbahor Abdullayeva. 2015. Uzbek paremiology and idioms in modern usage. *Uzbek Linguistic Studies*, 7(2):55–70.

Manzura Abjalova, Umida Rashidova, Sarvinoz Rasulova, and Sarvinoz Sharipova. 2025. Determination of stylistic features of idioms in uzbek language. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 583–587.

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

Nizami Bayramov. 2014. Azerbaijani idioms and their semantic features. *Azerbaijan Journal of Linguistics*, 12:88–102.

Hasan Biyik, Patrick Lee, and Anna Feldman. 2024. Turkish delights: a dataset on Turkish euphemisms. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 71–80, Bangkok, Thailand and Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas. Online version: `https://www.ethnologue.com/family/turkic/`.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Oguljaman Goçgeldiyeva. 2018. Phraseological units in turkmen and their cultural semantics. *Turkmen Studies*, 5(1):33–47.

Meriç Güven. 2020. *A Computational Analysis of Turkish Idioms*. Ph.D. thesis, Middle East Technical University.

Emrah Inan. 2025. Contrastive retrieval methodology for turkish metaphor detection and identification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(11).

Tolga Izdas, Omerhan Sancak, H. Toprak Kesgin, M. Kaan Yuce, and M. Fatih Amasyali. 2025. Turkish-e5: E5 model enhanced for turkish with multi-positive contrastive learning. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*.

L. Johanson and É.Á. Csató. 2015. *The Turkic Languages*. Routledge Language Family Series. Taylor & Francis.

Mehmet Karakus. 2020. Turkish idioms: A corpus-based semantic and structural study. *Journal of Turkish Linguistics*, 34(2):45–68.

Elena Mihail. 2020. Idioms and fixed expressions in modern gagauz. *Gagauz Philological Review*, 3(1):22–37.

Ali Püsküllüoğlu. 2012. *Türkçe Deyimler Sözlüğü*. Arkadaş Yayınevi.

Sh. Rahmatullayev. 1978. *O'zbek tilining izohli frazeologik lug'ati*. O'qituvchi, Toshkent.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Özge Umut, Atakan Site, Doğukan Arslan, and Gülşen Eryiğit. 2025. Exploring turkish idiomaticity with large language models. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 533–538.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zhengxuan Wu, Thanh-Son Nguyen, and Desmond Ong. 2020. Structured self-AttentionWeights encode semantics in sentiment analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 255–264, Online. Association for Computational Linguistics.

Hicran Yeşildere Aldan. 2020. Determination of common idioms in southwest (oghuz) turkish dialects. Master's thesis, Uşak University, Institute of Social Sciences, Uşak, Turkey. Advisor: Dr. Meriç Güven.

A Dataset Preparation for the Turkic Idiom Lexicon

Dataset Description. The Turkic Idiom Lexicon is a curated multilingual resource covering idiomatic expressions in four Turkic languages: Turkish (TR), Azerbaijani (AZ), Turkmen (TK), and Gagauz (GA). Each entry consists of an idiom surface form, a monolingual definition, and an example sentence illustrating idiomatic usage.

Source Data. The initial idiom inventory is derived from Yeşildere Aldan (2020), which provides manually aligned idiom lists across multiple Turkic varieties. While the source offers reliable idiom forms, many entries lack consistent definitions and contextualized examples.

Annotation and Normalization Pipeline. To enrich and standardize the lexicon, we employ a multi-stage normalization pipeline assisted by a large language model (GPT-4.1). For each idiom entry, the model generates: (i) a standardized idiom form, (ii) a monolingual definition, and (iii) an example sentence demonstrating idiomatic (non-literal) usage. All generations are performed in a zero-shot setting using fixed prompts to ensure consistency across languages.

> **System Instruction:** You are a linguist specializing in Turkic dialects (Turkey, Azerbaijan, Turkmen, Gagauz). Your task is to analyze the given list of idioms and generate meanings and example sentences for each. If an idiom cell in a dialect is empty or contains only a hyphen (-), leave those sections blank. Try to write the meanings according to the writing rules of the relevant dialect.
> **User Instruction:** Process the following JSON data. Each element is a set of statements. Desired Output Format (Return only this JSON structure):
> "TR-Deyim": "Input", "TR-Anlam": "Meaning in Turkish", "TR-Ornek": "Example Sentence"
>
> "AZ-Deyim": "Input", "AZ-Anlam": "Meaning in Azerbaijani Turkish", "AZ-Ornek": "Example Sentence",
>
> "TK-Deyim": "Input", "TK-Anlam": "Meaning in Turkmen Turkish", "TK-Ornek": "Example Sentence",
>
> "GA-Deyim": "Input", "GA-Anlam": "Meaning in Gagauz Turkish", "GA-Ornek": "Example Sentence"

Quality Control and Verification. Each generated entry is independently re-evaluated by a separate LLM-based verifier acting as a quality control model. The verifier assesses entries along three criteria:

- **Semantic Fidelity**: whether the generated definition accurately reflects the intended idiomatic meaning of the source expression,
- **Definition-Example Consistency**: whether the example sentence correctly instantiates the provided definition,
- **Idiomaticity**: whether the example reflects figurative rather than literal usage.

The verifier produces structured judgments (Accept, Reject) accompanied by confidence scores. Entries that fail any criterion, exhibit internal inconsistencies, or receive low confidence scores are automatically removed through deterministic, rule-based filtering.

LLM Prompt for Quality Control and Verification. Each generated idiom entry is independently evaluated using the following verification prompt:

> **System Instruction:** You are given an idiom, its definition, and an example sentence. Evaluate the entry according to the criteria below.
> **Evaluation Criteria:** 1. **Semantic Fidelity**: Does the definition accurately reflect the conventional figurative meaning of the idiom?
> 2. **Definition-Example Consistency**: Does the example sentence correctly instantiate the provided definition?
> 3. **Idiomaticity**: Is the idiom used figuratively in the example sentence rather than

literally?
For each criterion, answer with "Yes" or "No".
Finally, provide an overall decision:

- ACCEPT: if all criteria are satisfied
- REJECT: if at least one criterion is not satisfied

Return the output strictly in the following JSON format:

```
{
"semantic-fidelity": "Yes/No",
"definition-example-consistency": "Yes/No",
"idiomatic-usage": "Yes/No",
"final-decision": "Accept/Reject"
}
```

Filtering and Post-processing. Only entries labeled ACCEPT by the verifier are retained in the final dataset. Additional heuristic filters remove malformed outputs, incomplete entries, and surface-form duplicates. No manual post-editing is applied after filtering.

B Prompt Templates

The LLM prompts used for monolingual and cross-lingual semantic evaluation are provided here for reproducibility. For both settings, the LLM receives the idiom and the meaning and returns a JSON structure with two fields: *analysis* and *quality* (*HIGH*, *MEDIUM*, or *LOW*).

Monolingual Idiom-to-Meaning Evaluation Prompt. Idiom and meaning are provided in the same language.

System Instruction: You are a linguist specializing in figurative language and idioms. Your task is to evaluate whether the provided meaning accurately reflects the idiomatic usage in the given language. You must judge the semantic correctness and provide a brief justification.
User Instruction: Given the following idiom and meaning in the same language, evaluate whether the meaning accurately reflects the figurative idiomatic sense.

Idiom: "<IDIOM>"

Meaning: "<MEANING>"

Respond ONLY in the following JSON format:

```
{
"analysis": "short explanation",
"quality": "HIGH | MEDIUM | LOW"
}
```

Labeling Guidelines:
HIGH - The meaning fully reflects the figurative idiomatic sense.
MEDIUM - Partially aligned but missing nuance.
LOW - Literal, incorrect, or misleading.

Cross-Lingual Idiom-to-Meaning Evaluation Prompt. Idiom and meaning are provided in two different languages.

System Instruction: You are a linguist specializing in multilingual idioms across the Turkic language family. Your task is to assess whether the meaning provided in another language conveys the same figurative concept as the idiom in the source language.
User Instruction: Given the following source idiom and target language meaning, determine whether the meaning corresponds semantically to the figurative usage of the idiom.

Source Idiom (Language: <SRC_LANG>)
“<IDIOM>”

Target Meaning (Language: <TGT_LANG>)
“<MEANING>”

Respond ONLY in the following JSON format:

```
{
"analysis": "short explanation",
"quality": "HIGH | MEDIUM | LOW"
}
```

Interpretation guidelines::

- HIGH: meaning conveys equivalent figurative concept
- MEDIUM: partially aligned but culturally or semantically shifted
- LOW: unrelated or literal translation mismatch

C Idiom Span Detection Training Configuration

For Turkish (TR), each model is fine-tuned for 5 epochs using `AdamW`, batch size 8, learning rate 5×10^{-5}, and weight decay 0.01. Evaluation is performed at each epoch, and we enable `load_best_model_at_end` with F1 as the selection metric. Training is conducted on a single NVIDIA A100 GPU.

D Few-Shot Adaptation

Figure 4 presents the full learning curves for each encoder across the four target languages (AZ, TK, GA, UZ). These curves illustrate how performance evolves from zero-shot to full fine-tuning and provide complementary evidence for the effectiveness of few-shot adaptation.

E Idiom Semantic Retrieval Experimental Details

All semantic retrieval experiments were conducted on Google Colab Pro using a single NVIDIA A100 GPU. We use PyTorch 2.1, HuggingFace Transformers 4.44, and SentenceTransformers 3.0.

Training Procedure. All models are trained for one epoch (configurable) with a batch size of 16 and learning rate of 2×10^{-5}, using `AdamW` with weight decay 0.01. Early stopping is enabled, and the best checkpoint is selected by development-set NDCG. All evaluations are based on a single run per configuration due to computational constraints.

Loss Functions. SentenceTransformers models are trained using MultipleNegativesRankingLoss. HuggingFace encoder models use CosineEmbeddingLoss applied to pooled embeddings.

F Significance Direction of Monolingual Idiom-to-Meaning Retrieval (Task 2)

The Wilcoxon signed-rank analysis (Table 6) shows that language-specific fine-tuning yields statistically significant improvements for most models and pooling strategies, while ConvBERT-TR shows no significant change, indicating limited sensitivity to semantic supervision.

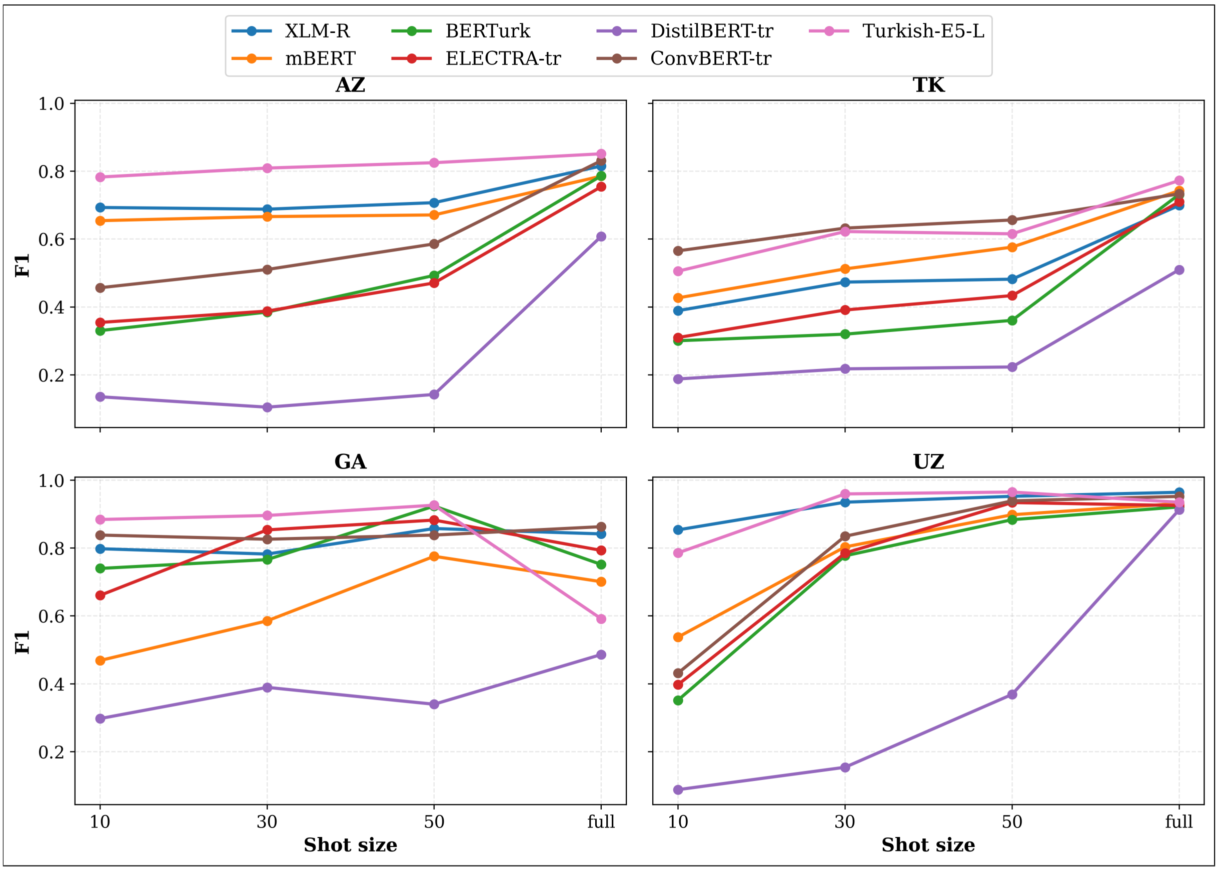

Figure 4: Few-shot adaptation results with increasing shot sizes (10/30/50/full) for each model and target language. Rapid early gains indicate that minimal supervision is sufficient to boost idiom span detection performance.

Model	Mean (Last)	CLS (Last)	Max (Last)	Mean (First+Last)
XLM-R	6.82e^{-02}	9.64e^{-01}	3.48e^{-01}	**3.53e**$^{-58}$
mBERT	**7.95e**$^{-03}$	**2.20e**$^{-04}$	**2.37e**$^{-05}$	**5.56e**$^{-10}$
BERTurk	**3.00e**$^{-93}$	**3.00e**$^{-93}$	**3.00e**$^{-93}$	**3.00e**$^{-93}$
ELECTRA-TR	**3.94e**$^{-05}$	**3.94e**$^{-05}$	**3.94e**$^{-05}$	**3.94e**$^{-05}$
DistilBERT-TR	**5.80e**$^{-08}$	**5.80e**$^{-08}$	**5.80e**$^{-08}$	**5.80e**$^{-08}$
ConvBERT-TR	6.67e^{-01}	6.67e^{-01}	6.67e^{-01}	6.67e^{-01}
Turkish-E5-L	**5.41e**$^{-139}$	**1.26e**$^{-127}$	**6.15e**$^{-132}$	**5.67e**$^{-139}$
Multilingual-E5	**1.41e**$^{-117}$	**7.69e**$^{-86}$	**2.39e**$^{-105}$	**7.44e**$^{-117}$
MPNet	**3.13e**$^{-122}$	**3.33e**$^{-114}$	**3.13e**$^{-101}$	**4.06e**$^{-14}$

Table 6: Wilcoxon signed-rank test (p-values) for Task 2 comparing pretrained vs language-specific fine-tuned models. Pooling strategies correspond to: *CLS (last layer)*, *mean pooling over tokens (last layer)*, *max pooling over tokens (last layer)*, and *mean pooling over tokens using the sum of first and last layers*. Bold values indicate statistical significance ($p < .05$).