SimIdioms: A Corpus and Benchmark for Ukrainian Idiom Translation

Yaryna Petruniv
Ukrainian Catholic University
petruniv@ucu.edu.ua

Iuliia Makogon
Independent Researcher
juogon@gmail.com

Roman Kyslyi
Kyiv School of Economics
rkyslyi@kse.org.ua

Abstract

We present a corpus of aligned Ukrainian–English idiomatic expressions and a comprehensive evaluation of six large language models on the task of translating sentences containing idioms. The corpus is constructed by linking entries across multiple phraseological dictionaries and the MIDAS corpus using vector similarity search, enriched with figurative meanings, contextual sentences from the UberText fiction corpus, and semantic transparency scores. We evaluate Gemini 2.5 Flash, Claude Haiku 4.5, Gemma 3 12B, Qwen3-30B-A3B, LapaLM, and Tiny Aya Global in both Ukrainian-to-English and English-to-Ukrainian directions under default and context-augmented prompting. Our evaluation of 65,723 translations reveals a pronounced direction asymmetry, with all models performing substantially worse when translating into Ukrainian. Identifying the source idiom in the prompt improves quality for most models in Ukrainian-to-English but has limited effect in the reverse direction, suggesting that the bottleneck there is morphological generation rather than idiom recognition. We additionally show that semantic transparency of idioms is only weakly correlated with translation quality. We release the corpus[1] and evaluation framework[2] to support research on idiomatic translation for mid-resource languages.

1 Introduction

Idiomatic expressions pose a fundamental challenge for machine translation: their meaning cannot be derived compositionally from constituent words, requiring models to recognize figurative usage and retrieve appropriate equivalents in the target language. While recent large language models have demonstrated impressive translation capabilities on standard benchmarks (Paniv, 2025), their ability to handle phraseological units remains underexplored, particularly for mid-resource languages such as Ukrainian.

This work addresses this gap with three contributions. First, we construct a corpus of aligned Ukrainian–English idiom pairs enriched with figurative meanings, contextual sentences, and semantic transparency scores (Section 3). Second, we evaluate six LLMs in both translation directions under default and context-augmented prompting, using an LLM-as-judge evaluation paradigm (Section 4). Third, we provide a detailed analysis of the factors influencing translation quality, including direction asymmetry, context utilization, semantic transparency, and model-specific failure modes (Sections 5–6).

2 Related Work

2.1 Multi-Word Expression and Idiom Corpora

Idioms are a subclass of multi-word expressions (MWEs) — conventionalized sequences of words whose meaning is often not fully derivable from their parts. The construction of idiom resources for computational research has received growing attention, evolving significantly in their creation methodologies from manual curation to AI-assisted generation. A prominent example of datasets that heavily rely on manual or crowdsourced annotations applied to automatically extracted text is the MAGPIE corpus (Haagsma et al., 2020), a single-language English resource containing tens of thousands of instances. To accelerate the development of such resources and handle their inherent complexity, researchers have increasingly adopted model-in-the-loop and LLM-generated approaches. The examples are FLUTE GPT-3 assisted dataset (Chakrabarty et al., 2022) or the recent MIDAS corpus (Kim et al., 2025) which employs LLMs for the initial refinement of idiomatic meanings followed

[1] https://huggingface.co/datasets/KSE-RESEARCH-Group/sim-idioms

[2] https://github.com/petrunivyaryna/sim-idioms

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 41–52

by native-speaker annotation.

These resources also differ fundamentally in their approach to multilinguality. While corpora like MAGPIE and FLUTE are strictly monolingual, the MIDAS corpus provides separate, non-parallel data for six typologically diverse languages. In contrast, datasets such as LIdioms (Moussallem et al., 2018) are explicitly designed as parallel corpora where data for different languages are manually synchronized and linked based on semantic equivalence. Flor et al. (2025) provide a comprehensive survey of these idiom datasets across psycholinguistic and computational paradigms, highlighting the persistent scarcity of high-quality resources for non-English languages. Furthermore, there remains a critical linguistic blind spot: none of these established benchmarks currently targets or supports the Ukrainian language.

2.2 Psycholinguistic Characteristics of Idioms

The psycholinguistic literature (Nunberg et al., 1994) classifies idioms as normally decomposable, abnormally decomposable, or semantically non-decomposable. Transparency is defined as the ease with which the structural motivation of an idiomatic expression can be deduced from its literal analysis. In our work, we operationalize this construct as the cosine similarity between idiom phrase embeddings and figurative meaning embeddings, following a distributional approach to transparency scoring. Computational assessment of semantic transparency was previously utilized by Gao et al. (2025) for Chinese idiom datasets. Kim et al. (2025) investigated whether LLMs rely on memorization or genuine reasoning when processing idiomatic expressions, finding that model performance degrades substantially on novel or low-frequency idioms. This motivates our evaluation design, which tests models on a diverse set of Ukrainian idioms with varying degrees of transparency.

2.3 Metrics for Idiom Translation

The inadequacy of surface-level metrics such as BLEU and METEOR for evaluating figurative language translation has been widely noted. Yang et al. (2025) demonstrated that reference-based metrics systematically underestimate the quality of idiomatic translations that use valid but lexically different target-language equivalents. Li et al. (2023) proposed augmenting translation systems with an idiom knowledge base (IdiomKB) and showed that standard metrics fail to capture improvements in figurative adequacy. These findings motivate our adoption of an LLM-as-judge evaluation paradigm (Zheng et al. 2023, Donthi et al. 2025), which can assess both meaning preservation and idiomatic adequacy without being constrained to exact lexical matches.

2.4 LLM Translation Benchmarking for Ukrainian

Several works address the need to evaluate LLM performance on tasks tailored to Ukrainian that may require cultural understanding. Recognizing limitations in translation capacity is especially important in adaptations of metrics constructed for high-resource languages with different grammar and morphology, e.g., Kravchenko et al. (2025), where translated cross-lingual pairs were used to assess LLM moral and cultural alignment. Researchers encounter difficulties applying traditional translation metrics to LLM-generated texts: for example, Paniv et al. (2025) report SacreBLEU on Multi30K-UK (Saichyshyna et al., 2023), a Ukrainian extension of the Multi30K multimodal benchmark. The recent LLM benchmark by Paniv (2025) includes translation evaluation on FLORES and LongFLORES for English–Ukrainian pairs with the BLEU metric. However, as demonstrated by Yang et al. (2025), such metrics do not suit phraseologically rich texts well, motivating the need for evaluation approaches specifically designed for figurative language.

3 Dataset Creation

3.1 Data Sources

We constructed the corpus of Ukrainian idioms and their English equivalents using multiple sources and matching strategies. We are grateful to the Condor Publishing House for granting permission to use the "Ukrainian-English and English-Ukrainian Phraseological Dictionary" (Horot et al., 2024) for academic purposes. Our dataset is primarily based on this dictionary, which provides Ukrainian and English idioms with corresponding translations but lacks consistent definitions or usage examples. To enrich the corpus with figurative meanings and contextual sentences, we incorporated the "Dictionary of Phraseological Units of the Ukrainian Language" by Bilonozhenko et al. 2003, which provides Ukrainian idioms with figurative interpretations and usage examples, as well as the English portion of the MIDAS corpus (Kim et al.,

Pattern	Count
VERB + obl	631
VERB + obj	558
NOUN + amod	251
NOUN + nmod	155
NOUN + case	145
VERB + obj + obl	115

Table 1: Most common syntactic patterns (root POS + dependency labels).

2025), which contains English idioms with figurative meanings and example sentences. Together, these resources allow the final corpus to capture not only cross-lingual idiom alignments, but also meaning and context on both sides.

We applied OCR to scanned dictionary pages that were not available in digital form.

3.2 Context for idioms retrieval

To address the shortage of contextual sentences for idioms, we used the UberText corpus (Chaplynskyi, 2023), focusing on its Fiction subset. This corpus contains modern Ukrainian texts segmented into sentences. The goal of this step was to automatically identify sentences in which a target idiom appears, including inflected forms and variations in word order within the idiom.

Each idiom can be parsed into a root token and its immediate dependency relations. We used this dependency representation to build reusable matching templates for idiom retrieval. For each idiom, we extracted the POS tag of the root token and the dependency labels of its direct children. We then grouped idioms into clusters that share the same root POS and the same set of dependency relations. This clustering step allowed us to define one dependency pattern per cluster, rather than writing a separate pattern for every idiom. Table 1 summarizes the most common syntactic patterns in our idiom set.

For each cluster, we constructed a dependency pattern using spaCy's dependency matcher [3]. The pattern requires a root token with a given POS tag and child tokens attached to the root with the required dependency labels. In parallel, we built a lookup map for each pattern that maps a tuple of lemmas (the root lemma and the required child lemmas) to the corresponding idiom.

[3] https://github.com/explosion/spacy-models/releases/tag/uk_core_news_md-3.8.0

We then applied the dependency matcher to sentences from the UberText Fiction subset. Each sentence was parsed and checked against the set of dependency patterns. When a pattern matched, we formed the same lemma tuple used in the idiom lookup map (the root lemma plus the lemmas of the required dependency children) and used it to retrieve the corresponding idiom for that pattern.

To address false positives after dependency matching, we added a validation step using Gemini 2.5 flash. For each idiom-sentence pair, the model verified the phrase presence and labelled the usage as idiomatic or literal.

This approach helped us obtain additional context for 1,639 Ukrainian idioms.

3.3 Vector Indexing in Qdrant

Because the three main resources provide different types of information, they are not aligned at the entry level (see Table 2). Condor provides Ukrainian–English pairs, whereas Bilonozhenko and MIDAS provide figurative meanings and example sentences for Ukrainian and English idioms, respectively. To merge these sources into a single corpus, we needed a way to identify corresponding idiom entries across resources. To this end, we used Qdrant, a vector database for similarity search, to index idiom texts as embeddings and retrieve semantically similar candidate entries across sources.

We first transformed all sources into a unified set of records. Each record corresponds to a text fragment taken from a dictionary entry. Depending on the source, this fragment is an idiom string, a translation equivalent, a figurative meaning, or an example sentence. Each record is stored together with metadata: a stable idiom identifier, the source name, and the language.

We embedded the text of each record using the multilingual sentence embedding model Multilingual-E5-large (Wang et al., 2024) and stored the resulting vectors in a Qdrant collection configured with cosine similarity. The embedding vector is stored in the index, while all metadata fields are stored as payload.

3.4 Building the Final Aligned Corpus

We treated corpus construction as a linkage problem across sources: entries that refer to the same idiom should be connected, and the final corpus should consist of consistent cross-source groups.

Source	# UK	# EN	# Meanings	# Examples	Notes
Condor (UK–EN)	4,565	12,159	✗	✗	EN are translation equivalents (not necessarily idioms)
Condor (EN–UK)	14,342	5,638	✗	✗	UK are translation equivalents (not necessarily idioms)
Bilonozhenko	3,304	✗	5,176	8,441	UA idioms
MIDAS (EN)	✗	12,662	11,806	19,410	EN idioms

Table 2: Summary statistics for the main sources used to construct the corpus.

First, we defined which source–field subsets were allowed to be linked. We generated links only for the pairings shown in Table 8.

The matching pipeline consists of several steps:

1. **Candidate retrieval:** For each pair, we used Qdrant to retrieve the most similar candidates from the target subset for every entry in the source subset.

2. **Threshold filtering:** We kept a candidate link only if its similarity score exceeded a predefined threshold. In our implementation, the threshold depends on the syntactic type of the idiom root in the source subset: 0.90 for common POS types and 0.95 for all others.

3. **Edge storage:** All accepted matches were stored in a separate Qdrant collection as edge records. Each edge stores the two connected idiom IDs, along with metadata such as the matching direction and the reason for the link.

After the matching step, we built an undirected graph in which the nodes correspond to the idiom identifiers and the edges correspond to the links from the link collection. We defined the final clusters as the connected components of this graph. Each connected component was assigned a unique cluster identifier (e.g. SIM-i) and represents a set of entries that are linked directly or via intermediate nodes.

Field	EN	UK
example	2,885	2,737
figurative meaning	1,550	363
figurative meaning translated	173	830
idiom	3,595	2,751
translation	3,427	3,026

Table 3: Item counts per field in the final corpus of 2,262 clusters.

To reduce manual workload, we assigned each cluster an initial decision label: KEEP or ANNOTATE. The label was determined based on source coverage. All ANNOTATE-labeled clusters were reviewed manually. The counts of unique entities grouped into 2,262 clusters are presented in Table 3.

3.5 Semantic Transparency Scoring

To characterize the difficulty of each idiom cluster, we computed semantic transparency scores that quantify how close an idiom's literal wording is to its figurative meaning. For each cluster, we embedded both the idiom phrase and its figurative meaning using Multilingual-E5-large (Wang et al., 2024) and computed the cosine similarity between the resulting vectors. Scores were computed separately for the Ukrainian and English sides, and an aggregate score was derived.

Out of 526 clusters with computed transparency scores, 77 were flagged as outliers, cases where the semantic transparency score was anomalously high, suggesting that the expression is more compositional than idiomatic. The remaining 449 clusters constitute the core evaluation set with genuine figurative expressions.

4 Experimental Setup

4.1 Models

We evaluated six large language models spanning different model families and sizes: Gemini 2.5 Flash (Google DeepMind, 2025), Claude Haiku 4.5 (Anthropic, 2025), Gemma 3 12B (Gemma Team, Google DeepMind, 2025), Qwen3-30B-A3B (Qwen Team, 2025), LapaLM (open large language model based on Gemma-3-12B adapted for Ukrainian language processing Team 2025), and Tiny Aya Global (Salamanca et al., 2026). The selection includes both proprietary API-based

models (Gemini, Claude) and open-weight models (Gemma, Qwen, LapaLM, Tiny Aya Global), allowing us to compare translation quality across different accessibility tiers. Notably, LapaLM is a Ukrainian-adapted model based on the Gemma 3 architecture, making the LapaLM–Gemma comparison a direct test of whether Ukrainian-specific fine-tuning improves idiomatic translation.

4.2 Translation Directions and Prompt Types

Each model was tested in two translation directions — Ukrainian-to-English (UK→EN) and English-to-Ukrainian (EN→UK) — under two prompt conditions:

- **Default:** The model receives only the source sentence and a translation instruction.
- **With context:** The prompt additionally identifies the source idiom by name. No figurative meaning or candidate equivalent is provided — only the idiom string itself, sufficient to signal that a non-literal interpretation is required.

This design isolates the contribution of idiom *recognition* from the contribution of richer semantic scaffolding such as figurative meanings or candidate target-language equivalents (Li et al., 2023; Donthi et al., 2025). It allows us to measure both baseline translation ability and the marginal benefit of explicit idiom identification.

4.3 Evaluation

We employed an LLM-as-judge evaluation paradigm (Zheng et al., 2023) using Gemini 2.5 Flash as the judge model. Each translation was evaluated on a 3-point scale:

- **Score 3:** The figurative meaning is preserved *and* the translation uses an idiomatic expression in the target language.
- **Score 2:** The figurative meaning is preserved but the translation is literal (no target-language idiom).
- **Score 1:** The figurative meaning is lost or severely distorted.

We frame Score 3 as a measure of *idiomatic retrieval competence*: the model's ability to recall a target-language idiom and produce it in context, rather than translation correctness in an absolute sense. A meaning-preserving but literal translation (Score 2) is not a translation failure; it is a distinct outcome that we analyze separately. We adopt this framing in both directions, including English-to-Ukrainian, because a model that reliably produces idiomatic output where appropriate demonstrates deeper phraseological knowledge of the target language than one that only produces literal renderings, even when both convey the same meaning. This is the capability we aim to assess.

In addition to the numeric score, the judge annotated each translation with binary labels for *meaning preservation* and *target idiom usage*, together with free-text reasoning. The total evaluation corpus comprises 65,723 annotated translations: 32,158 for UK→EN and 33,565 for EN→UK.

To validate the reliability of the LLM judge, we conducted a manual evaluation on a random sample of 100 sentences from the data set. These examples were independently annotated by a human evaluator using the same scoring criteria. The annotations were then compared with the scores produced by the LLM judge.

The comparison showed that the judge achieved an accuracy of 87% for the translation of the UK→EN and 78% for the translation of the EN→UK, indicating that the LLM-based evaluation is reasonably aligned with human judgment for the task of assessing idiom translation.

A potential limitation of this approach is the judge model's knowledge of Ukrainian idioms. Determining idiomaticity requires recognizing whether a phrase is an established Ukrainian idiom, which large language models may fail to do. To mitigate this issue, prompts include golden examples of acceptable translations and use separate instructions for each translation direction, with stricter idiomaticity guidance in the EN→UK setup.

5 Results

5.1 Overall Model Comparison

Table 4 and Figure 1 present the mean translation scores across all conditions. A clear three-tier structure emerges: Gemini leads with a mean score of 2.54, followed by Claude (2.32) and Gemma (2.29) in a middle tier, while LapaLM (2.15), Qwen (2.08), and Tiny Aya (1.89) form the lower tier.

5.2 Translation Direction Asymmetry

All models perform substantially worse on EN→UK than on UK→EN, as visualized in Figure 2. The largest directional gap is observed for Qwen (−0.57), while Gemini shows the most sta-

Model	Overall	UK→EN	EN→UK	Δ
Gemini	2.54	2.66	2.42	−0.24
Claude	2.32	2.56	2.09	−0.47
Gemma	2.29	2.49	2.09	−0.40
LapaLM	2.15	2.34	1.98	−0.37
Qwen	2.08	2.37	1.80	−0.57
Tiny Aya	1.89	2.11	1.68	−0.43

Table 4: Mean translation quality scores (1–3 scale). Δ shows the drop from UK→EN to EN→UK.

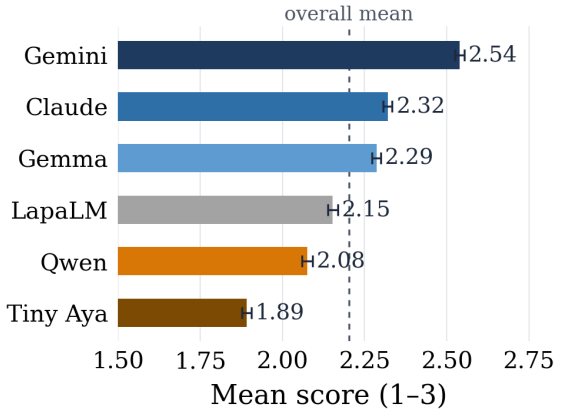

Figure 1: Mean translation quality scores by model across all conditions. Error bars show 95% bootstrap confidence intervals.

ble performance across directions (−0.24). This asymmetry is consistent with the morphological richness of Ukrainian and the greater challenge of generating idiomatic expressions in a morphologically complex target language.

In the UK→EN direction, the percentage of perfect translations (score 3) ranges from 40.4% (Tiny Aya) to 74.0% (Gemini). In the EN→UK direction, these rates drop substantially: from 17.0% (Tiny Aya) to 54.1% (Gemini). Meaning preservation rates remain comparatively stable across directions for the top models (Gemini: 91.8% vs. 93.7%), but degrade notably for lower-tier models (Tiny Aya: 70.6% vs. 61.2%).

5.3 Effect of Context

Table 5 and Figure 3 summarize the impact of identifying the source idiom in the prompt. In the UK→EN direction, context yields meaningful improvements for five out of six models, with the largest gain for Gemma (+0.233) and the smallest for LapaLM (+0.033).

In the EN→UK direction, context benefits are dramatically reduced. Only Gemini (+0.161) and Gemma (+0.133) show notable improvements. LapaLM, Qwen, and Tiny Aya show near-zero or negative deltas. This suggests that generating morphologically correct idiomatic Ukrainian is a bottleneck that additional semantic context alone cannot overcome.

LapaLM stands out as uniquely unable to leverage context in either direction ($\Delta_{\text{ctx}} = +0.033$ and $+0.000$), suggesting fundamental limitations in incorporating auxiliary information during generation.

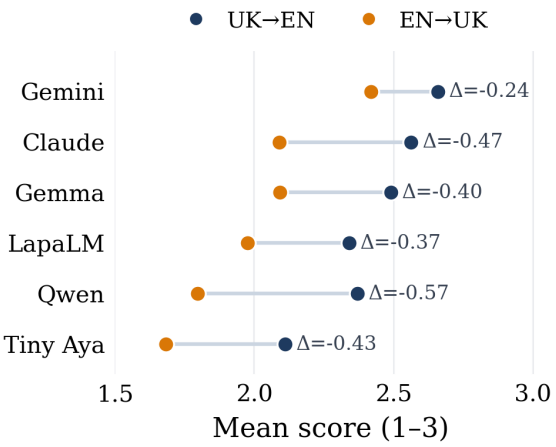

Figure 2: Direction asymmetry: performance drop from UK→EN to EN→UK per model.

	UK→EN		EN→UK	
Model	Default	Δ_{ctx}	Default	Δ_{ctx}
Gemini	2.587	+0.142	2.339	+0.161
Claude	2.454	+0.218	2.043	+0.094
Gemma	2.374	+0.233	2.026	+0.133
LapaLM	2.324	+0.033	1.975	+0.000
Qwen	2.278	+0.184	1.784	+0.024
Tiny Aya	2.062	+0.097	1.688	−0.009

Table 5: Default scores and context deltas (Δ_{ctx}) by direction. Positive values indicate improvement from context.

5.4 Score Distribution and Failure Modes

Table 6 details the score distribution for each model and direction. In UK→EN, Gemini achieves 74.0% perfect translations, whereas Tiny Aya reaches only 40.4%. In EN→UK, the proportion of complete failures (score 1) increases substantially for all models: Tiny Aya and Qwen produce over one-third score-1 translations.

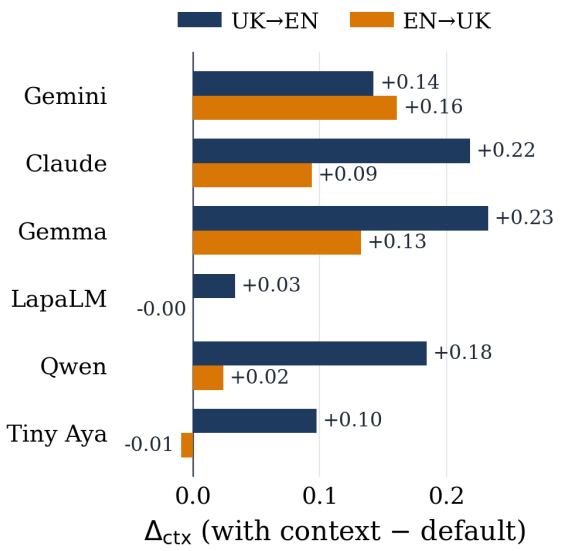

Figure 3: Context effect (Δ_{ctx}) by model and direction. Bars show the improvement from identifying the source idiom in the prompt.

Model	UK→EN (%) 1	2	3	EN→UK (%) 1	2	3
Gemini	8.2	17.8	74.0	12.2	33.8	54.1
Claude	12.8	18.2	69.1	23.1	44.9	32.0
Gemma	14.9	21.2	64.0	20.8	49.1	30.0
LapaLM	18.3	29.3	52.4	32.5	37.5	30.0
Qwen	18.4	26.2	55.4	39.2	42.1	18.8
Tiny Aya	29.4	30.2	40.4	48.6	34.4	17.0

Table 6: Score distribution by model and direction.

5.5 Semantic Transparency and Outlier Analysis

We examined whether semantic transparency — the degree to which an idiom's literal wording reflects its figurative meaning — predicts translation quality. Transparency scores were estimated automatically using sentence embedding similarity between idioms and their figurative meanings. For this analysis, we average translation scores across both prompt conditions to derive a single quality score per cluster. Of the 526 clusters with computed transparency scores, 466 have at least one translation in our evaluation set and are used in this analysis.

Figure 4 visualizes the relationship at the cluster level. We observe a weak but statistically significant positive correlation between semantic transparency and translation quality ($r = 0.136$, $p = 0.003$, $n = 466$). This indicates that more compositional expressions are marginally easier to translate, as expected.

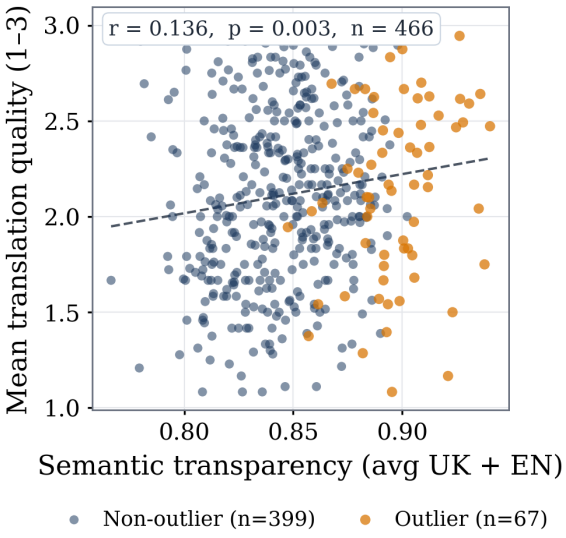

Figure 4: Semantic transparency vs. mean translation quality per cluster, averaged over both prompt conditions ($r = 0.136$, $p = 0.003$, $n = 466$). Each point is one idiom cluster.

Model	All	No outliers	Outlier only
Gemini	2.539	2.497	2.552
Claude	2.321	2.272	2.399
Gemma	2.286	2.236	2.389
LapaLM	2.153	2.066	2.295
Qwen	2.076	2.015	2.148
Tiny Aya	1.891	1.810	2.055

Table 7: Mean scores on clusters with transparency data: all clusters, non-outlier clusters, and outlier-only clusters (both prompts averaged).

Clusters flagged as outliers (77 of 526) — those with anomalously high semantic transparency — yield higher mean scores across all models (Table 7). However, removing these outlier clusters has minimal impact on overall scores (-0.04 to -0.09), confirming that the main findings are robust and not driven by quasi-compositional expressions.

6 Discussion

Our results reveal several important findings for the evaluation of idiomatic translation by LLMs.

Direction asymmetry. The pronounced performance gap between UK→EN and EN→UK demonstrates that translating *into* a morphologically rich language poses fundamentally different challenges. While models can often identify the figurative meaning of a Ukrainian idiom and render it in English, the reverse task — selecting and

correctly inflecting a Ukrainian idiom — proves substantially harder. The gap is largest for Qwen (-0.57) and smallest for Gemini (-0.24), suggesting that model-specific factors beyond language-pair difficulty are at play. We note, however, that morphological complexity and mid-resource pre-training exposure are observationally confounded for Ukrainian and we cannot fully separate their respective contributions; we discuss this further in the Limitations section.

Context utilization as a diagnostic. The differential ability to leverage context is diagnostic of model capabilities. Context helps most models in UK→EN but only Gemini and Gemma in EN→UK, suggesting that weaker models lack the morphological and syntactic competence to translate a known idiom into grammatically correct Ukrainian output. LapaLM is uniquely unable to leverage context in either direction ($\Delta_{\text{ctx}} = +0.033$ and $+0.000$), pointing to fundamental limitations in how this model incorporates auxiliary information during generation.

Ukrainian-specific fine-tuning does not help idiom translation. One of our most notable findings is that LapaLM — a model specifically adapted for Ukrainian based on the Gemma 3 architecture (Team, 2025) — performs *worse* than its base model Gemma across all conditions. In UK→EN, LapaLM scores 2.34 vs. Gemma's 2.49. In EN→UK, 1.98 vs. 2.09. The gap is even more pronounced with context: Gemma achieves context deltas of $+0.233$ (UK→EN) and $+0.133$ (EN→UK), while LapaLM's are $+0.033$ and $+0.000$. This may possibly happen due to catastrophic forgetting of pre-trained phraseological knowledge during Ukrainian-specific fine-tuning. This is consistent with observations by Paniv (2025), who found that instruction-tuned models sometimes underperform base models on Ukrainian tasks.

Model tiers and training data. The three-tier structure (Gemini > Claude/Gemma > LapaLM/Qwen/Tiny Aya) likely reflects differences in both pre-training data coverage and instruction-tuning quality for Ukrainian. API-based models with proprietary training pipelines (Gemini, Claude) generally outperform open-weight alternatives, consistent with findings on broader Ukrainian benchmarks (Paniv, 2025). The primary failure mode across all models is *literal translation*: models preserve semantic content but fail to produce target-language idioms, indicating that the challenge is figurative expression retrieval, not comprehension — a pattern also observed by Li et al. (2023) in multilingual settings.

Semantic transparency. The weak correlation between semantic transparency and translation quality ($r = 0.136$, $p = 0.003$) suggests that, while compositional idioms are marginally easier to translate, the primary difficulty lies in the model's phraseological competence rather than the transparency of individual expressions. This aligns with psycholinguistic findings that idiom processing involves both compositional and non-compositional pathways (Titone and Connine, 1999), and that even "transparent" idioms require culturally specific knowledge for appropriate translation.

7 Conclusion

We presented a comprehensive evaluation of six LLMs on the task of translating Ukrainian idiomatic expressions. Our corpus of aligned Ukrainian–English idiom pairs, enriched with figurative meanings, contextual sentences, and semantic transparency scores, provides a challenging testbed for assessing phraseological competence.

The evaluation of 65,723 translations reveals that:

1. All models perform substantially better on UK→EN than EN→UK, demonstrating that generating idiomatic Ukrainian is fundamentally harder than generating idiomatic English.

2. Identifying the source idiom in the prompt improves translation quality for most models in UK→EN but has limited or no effect in EN→UK, suggesting that morphological generation is the bottleneck.

3. Ukrainian-specific fine-tuning (LapaLM) does not improve idiom translation over the base model (Gemma), and in fact degrades both absolute performance and context utilization.

4. The primary failure mode is literal translation rather than meaning distortion: models generally preserve the intended meaning but fail to retrieve target-language idioms.

5. The model ranking is stable across conditions, with a clear three-tier structure: Gemini, Claude/Gemma, LapaLM/Qwen/Tiny Aya.

6. Semantic transparency has only a weak (though statistically significant) correlation with translation quality ($r = 0.136$, $p = 0.003$), and removing semantically transparent outlier clusters has limited impact on overall findings (-0.04 to -0.09).

These results highlight the need for evaluation frameworks that go beyond surface-level metrics when assessing figurative language translation. We release the corpus and evaluation code to support future research on idiomatic translation for Ukrainian and other mid-resource languages.

Limitations

Our study has several limitations. First, we evaluate a single language pair (Ukrainian–English); the findings may not generalize to other mid-resource languages with different morphological characteristics.

Second, we cannot fully disentangle Ukrainian's morphological complexity from its mid-resource status in pre-training data: most LLMs see substantially less Ukrainian than English, *and* Ukrainian morphology is considerably richer. Our claims about EN→UK morphological difficulty should therefore be read as joint claims about morphology and exposure; separating the two requires a controlled cross-lingual comparison, which we leave for future work.

Third, the LLM-as-judge evaluation, while scalable, may introduce systematic biases compared to human annotation. We validated judge outputs against expert annotations on a sample, but did not perform a qualitative error analysis. A related concern is that the same model family (Gemini 2.5 Flash) is used both as judge and as one of the evaluated translators, which could produce a self-preference bias; we mitigated this with golden references and explicit rubrics rather than open-ended preferences, and a fully independent judge would strengthen future evaluations.

Fourth, the corpus covers primarily literary and journalistic idioms from dictionaries and the UberText fiction subset — domain-specific idioms (e.g., legal, medical) are underrepresented, and coverage is bounded by dictionary publication dates.

Finally, transparency scores were computed with a single embedding model (Multilingual-E5-large), and our evaluation is limited to six models at a single point in time; different embeddings or future model versions may yield different results.

Ethics Statement

The idiom corpus was constructed from publicly available resources and a dictionary used with explicit permission from the Condor Publishing House. The UberText corpus is publicly available for research purposes. We used LLMs for evaluation, which may reflect biases present in their training data. The work focuses on evaluation rather than deployment, and we do not foresee direct negative societal impacts.

License and data release. The released corpus contains derived structured data (cluster identifiers, alignment graph, transparency scores, and contextual sentences retrieved from UberText) together with the English-side content from MIDAS (Kim et al., 2025). We make the following terms explicit:

- The English-side content and the cluster-graph structure are released under CC BY-SA 4.0.
- The Ukrainian idiom strings, figurative meanings, and usage examples derived from the "Ukrainian-English and English-Ukrainian Phraseological Dictionary" (Horot et al., 2024) and the "Dictionary of Phraseological Units of the Ukrainian Language" (Bilonozhenko et al., 2003) are included with the permission of the respective publishers and are provided strictly for academic research use. This portion of the corpus is *not* licensed under CC BY-SA 4.0. Users intending to use the Ukrainian portion for commercial purposes must contact the Condor Publishing House directly.
- Following the MIDAS release (Kim et al., 2025), any portion of the English-side content that originates from Wiktionary is governed by the CC BY-SA 4.0 license and inherits its requirements: attribution to the original Wiktionary entries, redistribution under the same license, and explicit indication of any modifications. This requirement applies even in academic redistribution.

We release the corpus and evaluation code to promote transparency and reproducibility.

Acknowledgements

We are grateful to the Condor Publishing House for granting permission to use the "Ukrainian-English and English-Ukrainian Phraseological Dictionary" for academic purposes.

References

Anthropic. 2025. Claude 4.5 Haiku.

V. M. Bilonozhenko, I. S. Hnatiuk, V. V. Diatchuk, N. M. Nerovnia, and T. O. Fedorenko. 2003. *Slovnyk frazeolohizmiv ukrainskoi movy [Dictionary of Phraseological Units of the Ukrainian Language]*. Naukova Dumka, Kyiv.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025)*.

Hui Gao, Jing Zhang, Peng Zhang, and Chang Yang. 2025. Consistency rating of semantic transparency: an evaluation method for metaphor competence in idiom understanding tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10460–10471, Abu Dhabi, UAE. Association for Computational Linguistics.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent AI model.

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.

Ye. I. Horot, Yu. V. Hromyk, L. K. Malimon, L. P. Pavlenko, A. B. Pavliuk, and O. O. Rohach. 2024. *Ukrainsko-anhliiskyi ta anhlo-ukrainskyi frazeolohichnyi slovnyk [Ukrainian-English and English-Ukrainian Phraseological Dictionary]*. Condor Publishing House, Kyiv, Ukraine.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.

Andrian Kravchenko, Yurii Paniv, and Nazarii Drushchak. 2025. UAlign: LLM alignment benchmark for the Ukrainian language. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 36–44, Vienna, Austria (online). Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Yurii Paniv. 2025. Isolating llm performance gains in pre-training versus instruction-tuning for mid-resource languages: The ukrainian benchmark study. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 876–883, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2025. Benchmarking multimodal models for Ukrainian language understanding across academic and cultural domains. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 14–26, Vienna, Austria (online). Association for Computational Linguistics.

Qwen Team. 2025. Qwen3.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Alejandro R. Salamanca et al. 2026. Tiny aya: Bridging scale and multilingual depth.

Lapa LLM Team. 2025. Lapa LLM v0.1.2-instruct: An efficient Ukrainian open-source language model. https://huggingface.co/lapa-llm/lapa-v0.1.2-instruct. Model based on Gemma-3-12B, developed by researchers from Ukrainian Catholic University, AGH University of Krakow, Igor Sikorsky Kyiv Polytechnic Institute, and Lviv Polytechnic.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655–1674.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025. Evaluating llms on chinese idiom translation.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Allowed Linkages

Table 8 lists the source–field subset pairings used during corpus construction.

Left subset		Right subset
MIDAS (field: idiom)	↔	Condor EN–UK (field: idiom)
MIDAS (field: idiom)	↔	Condor UK–EN (field: translation)
Condor UK–EN (field: idiom)	↔	Bilonozhenko (field: idiom)
Condor EN–UK (field: idiom)	↔	Condor UK–EN (field: translation)

Table 8: Allowed linkages between sources and fields.

B Model Details

Table 9 provides details of the six evaluated models. All open-weight models were served using vLLM with temperature 0.3. API-based models used their respective default settings.

Model	Size	Access	Notes
Gemini 2.5 Flash	N/A	API	Google DeepMind
Claude Haiku 4.5	N/A	API	Anthropic
Gemma 3	12B	Open	Google DeepMind
Qwen 3	30B	Open	Alibaba Cloud
LapaLM	12B	Open	UA-adapted
Tiny Aya Global	3.35B	Open	Cohere Labs

Table 9: Summary of evaluated models.

C Translation Prompt Templates

For the **default** condition, models received only the source sentence and a translation instruction:

> *Translate the following text from [Ukrainian/English] to [English/Ukrainian]. Output only the translation, nothing else.*
> *Input:* `{text}`

For the **context-augmented** condition, the prompt additionally identified the source idiom by name. Note that no figurative meaning, candidate translation, or example from the corpus was provided — only the idiom string itself, sufficient to signal that a non-literal interpretation is required while keeping the prompt minimal:

> *Translate the following text from [Ukrainian/English] to [English/Ukrainian]. This text is an example usage of the idiom "`{idiom}`". Use this context to produce an accurate and natural translation. Output only the translation, nothing else.*
> *Input:* `{text}`

D Evaluation Prompt Template (UK→EN)

LLM Judge Prompt (Ukrainian → English)

Role. You are an expert in idiom translation, specializing in Ukrainian idioms, their figurative meanings, and natural English idiomatic equivalents.

Task. Evaluate whether the English translation preserves the Ukrainian idiom's figurative meaning and whether it uses a natural English idiom or fixed expression.

Test data: Ukrainian sentence: `{src}` | Idiom: `{idiom}` | Meaning: `{meaning}` | Translation: `{tgt}` | Golden equivalents: `{gld}`

Judging procedure.

1. Identify the segment in the translation corresponding to the Ukrainian idiom.
2. Determine the intended figurative meaning: (a) use idiom meaning if present; (b) else infer from golden equivalents; (c) else infer from context.
3. Decide whether the translation preserves that meaning.
4. If preserved, decide whether the rendering is idiomatic: (a) idiomatic means a conventional English idiom or widely recognized fixed expression used by native speakers; (b) a plain descriptive paraphrase is not idiomatic; (c) if golden equivalents are provided, com-

pare the phrase from the translation against them for idiomatic equivalence; (d) if golden equivalents are empty, judge by general native-speaker usage. Use golden equivalents as the strongest signal; do not penalize for wording differences, but penalize if clearly less idiomatic than the golden equivalents.

Scoring rubric. **1 pt:** Figurative meaning not preserved. **2 pts:** Meaning preserved but expressed non-idiomatically. **3 pts:** Meaning preserved and expressed idiomatically.

Output: JSON with keys: `score` (1–3), `mapped_phrase_in_tgt` (string), `meaning_preserved` (bool), `uses_english_idiom` (bool), `reasoning` (2–5 sentences).

E Evaluation Prompt Template (EN→UK)

LLM Judge Prompt (English → Ukrainian)

Role. You are an expert in idiom translation, specializing in English idioms, their figurative meanings, and natural Ukrainian idiomatic equivalents.

Task. Evaluate whether the Ukrainian translation preserves the English idiom's figurative meaning and whether it uses a natural Ukrainian idiom with the same figurative meaning.

Test data: English sentence: `{src}` | Idiom: `{idiom}` | Meaning: `{meaning}` | Translation: `{tgt}` | Golden example: `{gld}`

Definition. An idiom is a group of words established by usage as having a meaning not deducible from those of the individual words.

Judging procedure.

1. Determine the intended figurative meaning using the idiom meaning and the sentence context.
2. Identify the segment in the translation corresponding to the English idiom.
3. Decide whether the translation preserves that figurative meaning.
4. If preserved, decide whether the rendering is idiomatic: (a) a word-for-word structural copy of the English idiom (calque) is NOT idiomatic; (b) a plain descriptive paraphrase is NOT idiomatic; (c) the phrase must be a fixed, conventional expression that native Ukrainian speakers would naturally produce and recognize; (d) if a golden example is provided and the mapped phrase differs from it, treat this as a strong signal the rendering is NOT idiomatic.

Scoring rubric. **1 pt:** Figurative meaning not preserved; includes mistranslation, literal translation, calque, or unnatural phrasing. **2 pts:** Meaning preserved and translation reads naturally, but no established Ukrainian idiom is used. **3 pts:** Meaning preserved, translation reads naturally, and uses an established Ukrainian idiom that native speakers would recognize and use.

Output: JSON with keys: `score` (1–3), `mapped_phrase_in_tgt` (string), `meaning_preserved` (bool), `uses_ukrainian_idiom` (bool), `reasoning` (2–5 sentences).

F Meaning Preservation Rates

Table 10 shows meaning preservation rates. Gemini achieves the highest rate in EN→UK (93.7%), even exceeding its UK→EN rate, suggesting its EN→UK failures are predominantly literal translations rather than meaning distortions.

Model	UK→EN	EN→UK
Gemini	91.8%	93.7%
Claude	87.2%	86.0%
Gemma	85.1%	85.3%
LapaLM	81.6%	77.3%
Qwen	81.6%	71.0%
Tiny Aya	70.6%	61.2%

Table 10: Meaning preservation rates by model and direction.

G LapaLM vs. Gemma: Detailed Comparison

Table 11 compares LapaLM directly with its base model Gemma 3. Despite Ukrainian-specific adaptation, LapaLM underperforms on every metric.

	Gemma		LapaLM	
Metric	**UK→EN**	**EN→UK**	**UK→EN**	**EN→UK**
Mean score	2.49	2.09	2.34	1.98
% Perfect (3)	64.0	30.0	52.4	30.0
% Failure (1)	14.9	20.8	18.3	32.5
Meaning pres.	85.1	85.3	81.6	77.3
Δ_{ctx}	+0.233	+0.133	+0.033	+0.000

Table 11: LapaLM vs. Gemma 3 (its base model).

H Outlier Analysis by Direction

Table 12 shows the outlier effect by direction. Outlier clusters (higher semantic transparency) are easier across both directions, with slightly larger effects in EN→UK.

	UK→EN		EN→UK	
Model	**No out.**	**Out. only**	**No out.**	**Out. only**
Gemini	2.617	2.666	2.393	2.486
Claude	2.534	2.662	2.036	2.245
Gemma	2.441	2.604	2.056	2.263
LapaLM	2.249	2.478	1.905	2.188
Qwen	2.293	2.582	1.770	1.894
Tiny Aya	2.003	2.319	1.640	1.901

Table 12: Scores on non-outlier vs. outlier-only clusters, by direction (both prompts averaged).