ID10M-JAM: Stress-Testing Idiom Identification Under Challenging Context

Kai Golan Hashiloni, Lior Livyatan, Ofri Hefetz, Alon Mannor, Bar Cohen, Kfir Bar
Data Science Institute, Reichman University, Herzliya, Israel
Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
kai.golanhashiloni@post.runi.ac.il,
{lior.livyatan, ofri.hefetz, alon.mannor, bar.cohen03}@post.runi.ac.il,
kfir.bar@runi.ac.il

Abstract

Large language models (LLMs) achieve strong performance on idiom identification benchmarks, yet their robustness to misleading contextual signals remains largely untested. We introduce ID10M-*JAM*, an adversarial extension of the ID10M dataset designed to *jam* model understanding by injecting coherent but conflicting context before each target sentence. For every sentence containing a potential idiomatic expression (PIE), we construct variants that deliberately invert contextual expectations: placing literal cues before idiomatic uses and idiomatic cues before literal ones. All variants are validated by human annotators to ensure naturalness and unambiguous interpretation for human readers. ID10M-JAM exposes systematic vulnerabilities in LLMs' contextual reasoning, pushing idiom identification to its breaking point.

1 Introduction

Idiomatic and other non-compositional expressions continue to pose a persistent challenge for computational models, despite the remarkable progress of large language models (LLMs). While idioms are only one subclass of multiword expressions (MWEs), they are unique in being both (1) formulaic, usually appearing in a fixed form, and (2) semantically non-compositional—the meaning of the whole cannot be fully inferred from the meanings of the parts. This non-compositionality is central not only to linguistic theory but also to practical natural language processing (NLP) applications. Misinterpreting idioms has been shown to lead to substantial downstream errors in machine translation (Baziotis et al., 2023; Dankers et al., 2022; Barreiro et al., 2013; Salton et al., 2014; Fadaee et al., 2018), spelling correction (Horbach et al., 2016), and semantic analysis (Cohen et al., 2022; Williams et al., 2015; Liu et al., 2017). While recent LLMs show improved idiom understanding compared to earlier systems, the extent to which these challenges persist in state-of-the-art models remains an open question. Idioms are fundamental to natural language use: English alone contains roughly 25,000 fixed expressions (Weinreich, 1969), comparable in scale to its core lexicon (Jackendoff, 1997). They occur frequently in everyday communication—every 3-4 minutes on average (Pollio et al., 1977)—amounting to an estimated 20 million uses over a speaker's lifetime (Cooper, 1998, 1999).

Original sentence from ID10M:
They asked me if we can *go dutch*.

Same sentence with enriched context (hard variant):
We were eating some German bread with Belgian chocolate when they asked me if we can *go dutch*.

Figure 1: Example of an original sentence from ID10M and a corresponding hard variant from ID10M-JAM. Highlighted are the potentially confusing spans.

Recently, Kim et al. (2025) released MI-DAS, a large-scale dataset of idioms in six languages along with their corresponding meaning. They demonstrated, through extensive experiments, that LLMs rely not only on memorization but also possess a more profound understanding grounded in identifying and reasoning about contextual cues. Recent work (Hashiloni et al., 2025) shows that LLMs can identify idiomatic expressions in running text using carefully designed prompts, sometimes outperforming supervised idiom detectors. A central challenge is that idioms can be surface-identical to literal expressions depending on context (e.g., "I spilled the beans in the kitchen"). Such expressions (in this example, "spill the beans") are termed potential idiomatic expressions (PIEs). The task, therefore, is to identify which PIEs in a document are used idiomatically. LLMs, especially in English, perform

Findings of the Association for Computational Linguistics: ACL 2026, pages 20846–20864

well on this task and use the surrounding context to infer whether a PIE is used idiomatically or literally (Hashiloni et al., 2025; Arslan et al., 2025; De Luca Fornaciari et al., 2024; Mi et al., 2025; Phelps et al., 2024). However, Gonen et al. (2025) demonstrated that contextual information can at times be harmful, because LLMs tend to learn spurious correlations among surrounding words, which can misguide their semantic understanding.

In this work, we examine how confusing context affects a model's ability to interpret PIEs and distinguish between their literal and figurative meanings. We introduce **ID10M-*JAM***, a new evaluation benchmark for idiom identification that probes generative LLMs under deliberately misleading contexts. Unlike standard setups with short, semantically clear samples, our benchmark imposes a substantially higher semantic load. Each context includes adversarial cues that are unambiguous to human readers yet bias toward the incorrect interpretation, requiring models to resolve literal versus figurative readings despite misleading surface signals. This design tests whether state-of-the-art LLMs are more error-prone than humans in contexts that humans find unambiguous.

Building on the ID10M dataset (Tedeschi et al., 2022), we modify sentences containing PIEs by enriching their surrounding context with such adversarial cues. We provide an example in Figure 1, where the PIE is originally used in its literal sense, but the added context introduces a figurative cue that could potentially mislead a model. Notably, despite the added signals, the intended literal meaning remains clear to human readers, allowing us to probe whether LLMs are confused in cases where human interpretation is unambiguous. This allows us to isolate model-specific susceptibility to contextual confusion. Using this data, we evaluate a range of LLMs—open-weight, closed-weight, and reasoning systems—under zero-shot and more sophisticated prompting setups that are effective for idiom identification.

Our contributions are threefold:

1. We introduce ID10M-JAM, the first benchmark targeting *confusing* PIE contexts, designed to evaluate LLMs' ability to distinguish between figurative and literal meanings in adversarial settings.

2. We provide a systematic evaluation of LLM performance on this task across English and German.

Dataset	Language
VNC-Tokens (Cook et al., 2008)	EN
Open-MWE (Hashimoto and Kawahara, 2009)	JA
Sporleder and Li (Sporleder and Li, 2009)	EN
IDIX (Sporleder et al., 2010)	EN
SemEval-2013 Task 5 (Korkontzelos et al., 2013)	EN
MAGPIE (Haagsma et al., 2020)	EN
EPIE (Saxena and Paul, 2020)	EN
AStitchInLanguageModels (Tayyar Madabushi et al., 2021)	EN, PT
Dodiom (Eryiğit et al., 2022)	IT, TR
ID10M$_{\text{gold}}$ (Tedeschi et al., 2022)	EN,DE, ES, IT
ID10M$_{\text{silver}}$ (Tedeschi et al., 2022)	EN,DE, ES, FR, IT, JA, NL, PL, PT, ZH
SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022)	EN, GL, PT
DICE (Mi et al., 2025)	EN
FLUID QA (Park et al., 2025)	EN, ZH, KR
ID10M-JAM (ours)	EN, DE

Table 1: Overview of various idiom datasets and the languages (ISO 639-1) they cover.

3. We draw general insights about how LLMs cope with adversarial cues in the context of figurative language understanding and identification.

All datasets, prompts, and evaluation code are publicly released under the CC BY-NC-SA 4.0 (datasets) and Apache-2.0 (code) licenses to ensure transparency and reproducibility.[1]

2 Related Work

2.1 Idiom Identification

MWEs pose a longstanding challenge in NLP due to their syntactic irregularities and non-compositional semantics (Constant et al., 2017). Early work relied on rule-based or statistical methods (Cook et al., 2007; Fazly et al., 2009; Shutova et al., 2010), later incorporating distributional semantics (Gharbieh et al., 2016; Nedumpozhimana and Kelleher, 2021). Transformer-based models with contextualized representations became dominant, with joint MWE-syntax architectures (Taslimipoor et al., 2020; Savary et al., 2023), rules-neural methods (Tanner and Hoffman, 2023), and sequence labeling models (Zeng and Bhat, 2021; Tedeschi et al., 2022; Hadj Mohamed et al., 2024).

Idioms form a subclass of MWEs whose meanings cannot be compositionally inferred (Timothy Baldwin, 2010). Idiom *identification* requires detecting idiomatic spans in text, in contrast to idiom *classification*, where candidate expressions are given. Neural approaches have been applied to both settings (Briskilal and Subalalitha, 2022; He

[1] https://github.com/Intellexus-DSI/ID10M-JAM

et al., 2024). More recently, LLMs have demonstrated strong performance on idiom classification and identification tasks (Arslan et al., 2025; De Luca Fornaciari et al., 2024; Mi et al., 2025; Phelps et al., 2024; Hashiloni et al., 2025), often surpassing fine-tuned models.

2.2 Idiom Classification and Identification Datasets

See Table 1 for an overview of existing datasets. Most idiom datasets focus on English and target classification rather than identification. Early resources such as VNC-Tokens (Cook et al., 2008), IDIX (Sporleder et al., 2010), and SemEval-2013 Task 5 (Korkontzelos et al., 2013) are influential but limited in scale and coverage. More recent datasets expand size or scope, including EPIE (Saxena and Paul, 2020), SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022), and MAGPIE (Haagsma et al., 2020), the latter directly supporting identification.

Multilingual resources include PARSEME (Ramisch et al., 2020; Savary et al., 2023) and AlphaMWE (Han et al., 2020), which provide broader MWE coverage. ID10M (Tedeschi et al., 2022) is the most prominent multilingual idiom identification benchmark, while Dodiom (Eryiğit et al., 2022) and AStitchInLanguageModels (Tayyar Madabushi et al., 2021) cover additional languages and settings. FLUID QA (Park et al., 2025) evaluates figurative language in multilingual dialogue contexts.

The most closely related work is DICE (Mi et al., 2025), which contrasts literal and figurative uses of idiomatic expressions while holding surface form constant, revealing substantial LLM failures. In contrast, we introduce adversarial contextual probes that bias interpretation, enabling a more targeted evaluation of robustness under semantically misleading yet human-resolvable conditions.

2.3 Synthetic Data for Downstream Tasks

The generation of high-quality training data remains a significant bottleneck in NLP, as data collection and annotation are expensive, time-consuming, or require specialized expertise. Recently, LLMs have emerged as powerful data generators, capable of producing synthetic examples that can substantially augment human-annotated data and support dataset construction. This is showcased in knowledge distillation with synthetic paired data (Lee et al., 2024), prompt-based generation with quality refinement pipelines (Nadăş et al., 2025), and task-specific prompt engineering (Li et al., 2023).

Arslan et al. (2025) used GPT-4 to generate multilingual idiom instances, showing that synthetic data, while weaker than human annotations, offers a good efficacy-cost trade-off when properly validated. We adopt a similar approach for generating our hard variants, combined with rigorous human filtering and refinement.

2.4 Adversarial Trigger Tokens

Recent work has explored how adversarial inputs can disrupt LLMs' predictions. Universal adversarial triggers—short token sequences that reliably flip model predictions when appended to inputs—have been studied extensively (Wallace et al., 2021), with subsequent work analyzing why such triggers are effective and how they behave geometrically in representation space (Subhash et al., 2023). Classical adversarial attacks typically use minimal token-level substitutions (Jin et al., 2020) or optimization-based perturbations (Zhao et al., 2022), while other approaches employ contextualized perturbations that maintain fluency while misleading models (Li et al., 2021). More broadly, sentence-level adversarial contexts, including methods that add misleading clauses or contextual cues while preserving coherence, have been studied in the context of NLP robustness (Goyal et al., 2023). Our work extends this line of research by introducing a dataset of systematically constructed adversarial contexts, in which misleading prefixes are added to sentences containing PIEs. This dataset is designed to assess whether models can distinguish figurative from literal language use when presented with adversarial cues that do not introduce ambiguity for human readers.

3 ID10M-JAM

3.1 Problem Formulation

Traditionally, the task of identifying idiomatic expressions is framed as a sequence labeling problem (Tedeschi et al., 2022; Ide et al., 2025; Hashiloni et al., 2025). Given an input sentence, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, where x_i represents an individual token, the objective is to generate an output label sequence, $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. Each label, y_i, is drawn from the set $\{\texttt{B-IDIOM}, \texttt{I-IDIOM}, \texttt{O}\}$, which collectively implements the standard BIO tagging scheme. This scheme indicates whether a token is at the beginning, inside, or outside of

an idiomatic expression. This approach inherently combines the challenges of expression identification and semantic disambiguation. The model must not only pinpoint the boundaries of an idiom but also correctly determine if the usage is figurative or literal, without being provided with any predefined list of candidates. This setup closely mirrors a real-world scenario.

3.2 Original Samples Filtering

We begin by manually filtering out unsuitable samples from the original ID10M dataset before generating the ID10M-JAM variants. Details on the annotation protocol and annotators are provided in Section 3.4.

During validation, we identified and addressed several types of issues. Some sentences exhibited annotation ambiguity that could compromise evaluation. For instance, "I thought I taught you how to *get over* a fence" and "Hi there, babe, *what's cooking*?" can be interpreted either literally or figuratively. Other cases involved incorrect labels: "Lemon Drop sind sehr ertragreiche Chilipflanzen...über 100 *Früchte tragen*" ("Lemon Drop chili plants are very high-yielding...produce over 100 fruits in a year") was mistakenly labeled as idiomatic despite a clearly literal usage of "Früchte tragen" ("bear fruit"), while "After passing the exam, I was *on cloud nine*" was incorrectly marked as literal.

We also observe additional issues, including sentences containing multiple PIEs beyond the annotated one, unnecessary English translations appended to German sentences, and minor typographical or punctuation errors. Ambiguous cases were discarded, whereas correctable issues were resolved and the sentences retained.

After filtering, the dataset comprises 191 English sentences and 155 German sentences. Further filtering details are provided in Section 3.4.

3.3 Automatic Data Generation

Dataset definition. For each original sentence, we define adversarially enriched variants $(v_1, \ldots, v_n)$ and denote them "hard variants". Note that this is a general term for all variants, not a graded notion of difficulty. In our formulation, each $v_i = c_i \oplus s$, where c_i is an adversarial context intended to confuse models at inference time, s is the original sentence, and $\oplus$ is the concatenation operator. The added context c_i introduces lexical cues pointing toward the **opposite** interpretation of the PIE in the original s. Specifically, c_i must follow the following requirements: (1) if p is used idiomatically in s, then c_i should evoke its literal meaning, and vice versa; (2) c_i must not create ambiguity for a human reader—the intended usage of p in s must remain unmistakable; and, (3) c_i may not introduce new PIEs, nor repeat p. For an illustration, see Figure 1.

Dataset	Language	Train	Test	
		# Sentences	# Sentences	# Variants
ID10M (Tedeschi et al., 2022)	EN	37,919	200	-
	DE	24,126	200	-
ID10M-JAM (Ours)	EN	-	178	534
	DE	-	137	411

Table 2: Overview of ID10M (original sentences) and our new ID10M-JAM datasets, where instances are hard variants generated in this work.

Dataset generation. We prompt Gemini 2.5 Pro[2] (Comanici et al., 2025) to generate *five* different variants from a single original sample in a single call, instructing it to maintain diversity and to use the default temperature of 1.0. We then prompt the same model a second time, assigning it the role of a validator and instructing it to identify and correct errors, replace invalid variants, and perform related validation steps. See cost details in Appendix A. As shown below, the LLM-based validator proved effective compared to human validation, substantially reducing both the proportion of invalid samples and the workload for human validators. The full prompts for both steps are provided in the project repository[3] for reproducibility and are based on the annotation guidelines given to human annotators as described below. To work with Gemini and some other models, we use the *agno*[4] framework, which provides convenient, standardized, and fully reproducible API calls across models and experiments, while abstracting and unifying the API designs of different providers.

Each variant is created with BIO annotations, where the original PIE is labeled as idiomatic only if it was figurative in the original sentence, and all remaining tokens are labeled "O", since no additional PIEs are introduced during context augmentation. We refer to each newly generated variant as a *hard* variant.

[2]This model was chosen after experimenting with other potential LLMs.

[3]Given their length, we omit them from the paper.

[4]https://github.com/agno-agi/agno

3.4 Human Validation

To ensure the quality of the automatically generated hard variants, we conducted human validation following a detailed annotation protocol, which is publicly available in the project repository and summarized below. Annotation was performed using the Label Studio[5] platform. See Figure 5 (Appendix E) for an annotation example and Appendix B for more information about the human annotators. Annotators were given (i) the original sentence s containing a PIE p and its gold label (figurative or literal), and (ii) a single generated variant v_i. Their task was to verify that v_i satisfies all validity constraints and preserves the intended label of p. Access to the original labels was intentional, as annotators were instructed to ensure that the generated context preserved the original PIE label and remained unambiguous. To clarify, their task was validation and correction, not curating from scratch. To mitigate LLM-generated stylistic biases, annotators actively revised instances that appeared unnatural or overly explicit. This human-in-the-loop refinement ensures that the adversarial cues remain as coherent and realistic as possible while preserving their ability to mislead models without confusing human readers. A hard variant is considered valid if it meets the following criteria: (1) the original sentence appears verbatim in the variant (with only minor punctuation differences permitted); (2) the variant is fluent, coherent, and fully interpretable to a human reader; (3) the added context introduces no new idioms and does not repeat p; (4) the original figurative or literal usage of p remains unambiguous for a human reader; (5) the added context contains new words or expressions that are semantically related to the opposite meaning of p and may steer the model away from its correct interpretation in s.

Annotators had three possible actions: leave the variant unchanged and mark it as valid; provide a minimally edited corrected version if the issue was fixable; or mark it as invalid if it violates the guidelines or could not be repaired. In problematic cases, annotators could optionally provide notes explaining the issue.

The authors reviewed all complex cases and resolved inconsistencies through collaborative discussion with the annotators.

Validation results. For the final dataset, we retain only original sentences with at least three valid variants and select three of them—this ensures sufficient variant coverage and balanced sampling for reliable evaluation. For example, for the original sentence "Is soaking *in hot water* good for us?", no variant proved to be valid, so it is not taken further. During this step, a couple of additional sentences proved too ambiguous and were discarded.

This process results in 178 originals corresponding to 534 variants in English and 137 originals corresponding to 411 variants in German. We release the corrected, filtered version of ID10M, together with detailed annotation documentation and ID10M-JAM, to our project's repository. For basic statistics about our dataset and ID10M, refer to Table 2.[6] ID10M-JAM contains 153 unique English PIEs and 92 German ones.[7]

4 Experiments

4.1 Evaluation Method

Standard idiom identification is typically evaluated using token-level F1 scores under the BIO tagging scheme (Tedeschi et al., 2022; Hashiloni et al., 2025), a metric that jointly reflects boundary errors (e.g., incorrect span extents) and semantic errors (e.g., misclassifying literal versus figurative usage). We aim to quantify the model's susceptibility to adversarial contexts, shifting the focus away from traditional evaluation set-ups. Our core evaluation criterion is defined as follows: Given a target PIE p that is correctly identified in its original sentence s, we test whether this detection is preserved when the model is presented with the corresponding hard variant v. A prediction is considered successful if the model correctly identifies p, regardless of its potential predictions for any other spans in the sentence. For decoder-only models that generate free-form text, predicted spans may differ from the canonical form of p, so we verify predictions using normalization and regular-expression matching.

To assess how a model's ability to classify PIEs as figurative or literal changes under adversarial context, we define two key metrics. First, S_M (success) denotes the number of hard variants whose *original* sentences were correctly classified by model M—that is, cases where p was identified correctly in the original ID10M sentence. Importantly, S_M is determined solely by the model's

[5] https://labelstud.io/

[6] Note that the ID10M dataset includes sentences in additional languages that are not shown in this table.

[7] We do not report comparable counts for ID10M, as the dataset was modified through the corrections described in Section 3.2.

		English		German	
Setting	**Model**	S_M	$ND_M(\%)\downarrow$	S_M	$ND_M(\%)\downarrow$
`Zero-shot`	GPT-4o mini	471.00±7.94	5.16±0.95	257.00±4.58	5.84±0.43
	Qwen2.5-72B	452.00±1.73	1.70±0.34	361.00±1.73	9.80±5.11
	Llama 4 Scout	481.00±6.93	6.30±0.76	292.00±6.24	8.89±1.43
	GPT-4o	483.00	6.42	333.00	12.01
	Claude 4 Sonnet	486.00	6.58	369.00	2.98
	Gemini 2.5 Flash	492.00±5.20	9.08±0.75	295.00±1.73	9.60±0.73
	Gemini 2.5 Pro	501.00	7.78	381.00	3.15
`Few-shot +SC+CoTBest`	GPT-4o mini	483.00±13.75	6.19±1.15	321.00±9.00	11.30±1.08
	Qwen2.5-72B	458.00±1.73	0.95±0.26	380.00±1.73	6.76±1.10
	Llama 4 Scout	487.00±3.46	6.36±0.31	308.00±13.86	8.74±1.00
	GPT-4o	495.00	6.06	390.00	8.72
	Claude 4 Sonnet	498.00	8.43	393.00	7.12
	Gemini 2.5 Flash	484.00±4.58	8.27±1.02	357.00±3.00	7.75±1.19
	Gemini 2.5 Pro	504.00	10.12	390.00	4.87
`Reasoning LLMs`	DeepSeek-R1	462.00	3.90	378.00	4.76
	o3-mini	477.00	5.45	324.00	8.95
`Encoders`	mBERT	298.80±6.91	23.96±3.66	206.40±20.28	19.78±1.86
	BERT	357.00±6.00	11.88±0.50	-	-
	GBERT	-	-	313.20±3.42	9.18±1.36

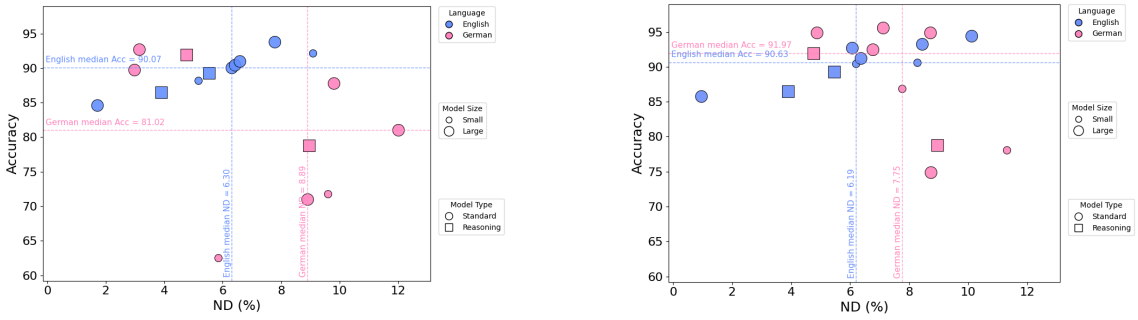
Table 3: Results on the ID10M-JAM dataset. S_M denotes the success count (out of 411 for German and 534 for English), and ND_M denotes negative drift, computed as described in Section 4.1. Lower ND_M indicates greater robustness (↓). Gemini models are reported separately for fairness, as they were involved in data generation. SC = Self-Consistency; CoT = Chain-of-Thought. Standard deviations, when applicable, are shown after ±.

performance on original sentences, independent of its predictions on variants. Second, F_M (flipped) counts how many of these correctly-classified originals experience a prediction flip in their variants—that is, cases where the model switches from correct on the original to incorrect on the variant (either predicting p as figurative when used literally, or as literal when used figuratively). By definition, $F_M \leq S_M$, since only originally correct predictions can flip. We define the *negative drift* as the proportion of such flips: $ND_M = \frac{F_M}{S_M}$. A higher ND_M indicates greater susceptibility to adversarial context.

Importantly, ND_M measures *robustness to adversarial context*, not overall performance. Models with lower original accuracy have fewer correctly-classified examples on which drift can occur, potentially yielding lower ND_M despite weaker understanding. Conversely, high-accuracy models may exhibit substantial drift. Thus, correct identification and adversarial resilience are not necessarily coupled.

4.2 Prompting LLMs

Since our goal is to evaluate LLMs on idiom identification, we explicitly prompt them to perform this task, following a growing line of work that employs LLMs for non-generative extraction tasks (Liu et al., 2023; Sun et al., 2023; Smădu et al., 2024; Hashiloni et al., 2025). For more details about the system prompt please refer to Figure 4 in Appendix C. We follow the prompting strategy introduced by Hashiloni et al. (2025) and evaluate each model under two configurations: a simple `zero-shot` setup and a more sophisticated prompt, which was shown to be the best-performing configuration in (Hashiloni et al., 2025) on the ID10M dataset. In the `zero-shot` configuration, the model is instructed to list the idioms present in a given sentence. No explanations are required, and no examples are provided. The temperature is set to 0.3 to ensure prediction stability. The second configuration, denoted `Few-shot+SC+CoTBest` in (Hashiloni et al., 2025), combines few-shot prompting, self-consistency, and chain-of-thought (CoT). The model is shown ten input-output examples (five with idioms and five without), randomly sampled from the ID10M training set. We apply self-consistency (`SC`; Wang et al. 2023) by sampling the model $n = 5$ times at temperature 0.8, to encourage output diversity, and retaining only idioms that appear in at least three outputs. We further use the `CoTBest` setup of Hashiloni et al. (2025), in which the model enumerates candidate PIEs with explanations and then selects at most one idiom—the one it judges most confidently idiomatic in context.



(a) `Zero-shot` (b) `Few-shot+SC+CoTBest`

Figure 2: Accuracy on the original ID10M dataset (y axis) vs. ND_M as detailed in Section 4.1 (x axis). SC = Self-Consistency, CoT = Chain-of-Thought. Encoders are excluded from this plot.

4.3 Models

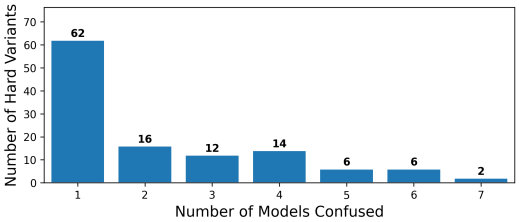
We evaluate a diverse set of models; checkpoint details and licenses are listed in Table 7 (Appendix I). We evaluate Google's Gemini 2.5 Flash and Gemini 2.5 Pro (Comanici et al., 2025), as well as OpenAI's GPT-4o as a strong closed-weight model, along with its smaller, more cost-efficient variant, GPT-4o mini. Similarly, we evaluate Anthropic's Claude 4 Sonnet. For comparison with open-weight models, we evaluate Llama 4 Scout and Qwen2.5-72B (Yang et al., 2025). We additionally evaluate two reasoning models: o3-mini and DeepSeek-R1 (DeepSeek-AI et al., 2025). These models are evaluated only in `zero-shot` setting, without sophisticated prompts, to isolate their built-in reasoning capabilities.

Fine-tuning encoder-based models. Alongside the LLMs, we fine-tune a set of encoder-based models on the complete training splits of the original ID10M dataset, using them as supervised baselines tailored directly to the task. Specifically, we train multilingual BERT (mBERT) (Devlin et al., 2019) as a cross-lingual model, as well as language-specific variants: BERT (Devlin et al., 2019) for English and GBERT (Chan et al., 2020) for German. To ensure robustness, each training and evaluation cycle is repeated five times with different random seeds. Further details on resources and hyperparameters are provided in Appendix D.

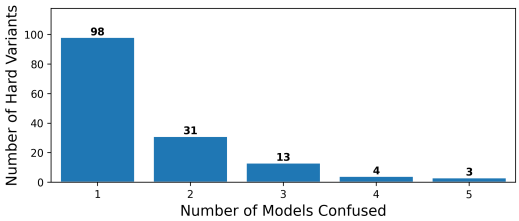
5 Results and Discussion

Across all experiments, we made approximately 200k API calls, at a total cost of about $280. To balance robustness and cost, most experiments use efficient models (GPT-4o mini, Llama 4 Scout, and Qwen2.5-72B), each evaluated with three seeds and reported as mean and standard deviation. More expensive models (e.g., GPT-4o and Gemini 2.5 Pro) are evaluated once per prompt configuration, while reasoning models (o3-mini and DeepSeek-R1) are used only once, under `zero-shot` prompting. Table 3 summarizes the results of all our experiments on the ID10M-JAM dataset. For each model and configuration, we report the corresponding S_M and ND_M. Detailed results, broken down by literal and idiomatic classes, are provided in Table 5 (Appendix F). Figure 2 illustrates the relationship between accuracy and ND_M under both prompting configurations, where each point corresponds to a single model, with shape, color, and size encoding reasoning capability, language, and model size, respectively. Accuracy is defined as the proportion of original ID10M sentences that are correctly classified (see Table 2 for the total counts per language). As the task requires the model to identify the PIE only when it is used figuratively, we consider a model that predicts no idioms for any input to represent the minimum accuracy. The English split contains 20/178 literal samples (11.24%) and the German one has 8/137 (5.8%). This distribution is similar to the original ID10M test split (20% and 9% literal, respectively), with the slight decrease due to our removal or correction of some original sentences, as detailed in Section 3. Figure 2 provides a convenient way to jointly consider both metrics, accuracy and ND_M. Models located in the upper-left corner are considered preferable, as they combine high accuracy with low susceptibility to adversarial contexts. Overall, we observe no clear relationship between model characteristics and robustness to adversarial context, making it difficult to predict which models are more likely to fail in this setting. We now discuss several aspects of the results in more detail.

Language comparison. Across nearly all set-



(a) English



(b) German

Figure 3: Histogram of the number of models that experienced a negative drift per hard variant. Encoders are excluded from this calculation and plot.

tings, German emerges as the more challenging language, consistent with the results reported by Hashiloni et al. (2025). For example, Figure 2 shows that the English median accuracy (90.97%) substantially exceeds German (81.02%), and German exhibits higher median ND_M (8.94% vs. 6.19%). We also see that the English models are slightly more pulled towards the upper-left corner than the German ones. This suggests that German PIEs introduce greater variability, or that models are generally better at handling English, unsurprisingly.

Figure 2 also reveals an opposite pattern across languages. In English, accuracy and ND_M show a slight positive correlation: better models solve harder samples yet are not necessarily more robust to adversarial context. In German, the trend reverses: a higher ND_M is associated with lower accuracy, suggesting that weaker models rely more heavily on surface cues and are therefore more susceptible to adversarial manipulation on the original task.

Prompt types. The `Few-shot+SC+CoTBest` prompt substantially improves accuracy for most models: for instance, GPT-4o German S_M rises from 333 to 390. We see an exception for Gemini 2.5 Flash in English, which decreases slightly from 492 to 484. Under these settings, German median accuracy (Figure 2) reaches 91.97%, compared to 81.02% in Zero shot, even surpassing the one in English (90.63%). German nonetheless retains a higher median ND_M (7.75% vs. 6.19%), indicating that robustness to adversarial context remains harder to achieve in German. Despite these accuracy gains, the prompt only slightly relaxes ND_M on average, especially in German (7.75% vs. 6.19%), with some models becoming more robust and others becoming less robust. This suggests that structured reasoning behaves unpredictably in the presence of adversarial cues.

Model size and scaling. Scaling shows a non-monotonic trend: larger models generally achieve higher accuracy, reflecting stronger semantic and figurative understanding, but size does not guarantee lower negative drift. Some strong models remain vulnerable (e.g., GPT-4o in German), while mid-sized models (e.g., GPT-4o mini) achieve competitive ND_M. Overall, higher accuracy does not imply robustness to adversarial contexts.

Reasoning vs. non-reasoning models. Overall, reasoning-oriented models tend to be only marginally more robust than their non-reasoning counterparts, in most cases. As shown in Table 3, DeepSeek-R1 seems to be slightly more robust than o3-mini with 3.85 vs. 5.45 in English and 4.76 vs. 8.95 in German.

Hard variants difficulty levels. Figure 3 shows the distribution of variant difficulty across the dataset. To estimate difficulty, we count how many LLMs (excluding encoder-based models) fail on each variant after correctly handling the original sentence; variants that confuse no models are omitted. Overall, 118 English and 149 German hard variants confuse at least one model (i.e., the sum of all bars in each chart), with most affecting only a small subset. This distribution reveals that our dataset contains a *range of challenge levels* rather than uniformly difficult examples: some variants prove challenging to nearly all models, while others expose vulnerabilities in only specific systems.

We analyze per-sentence confusion patterns in Table 6 and Figure 6 (Appendix F), and find varied distributions: some sentences yield many error-inducing variants, while others yield few or none. In English, models are more uniformly affected, typically failing on either all variants or none, whereas in German they are more sensitive to specific phrasings.

Encoders. We find that encoder models perform worse than LLMs on the basic identification

task, despite being fine-tuned for it. Notably, the language-specific encoder consistently outperforms the multilingual encoder on the original sentences and substantially reduces the negative drift.

Summary of key observations. (i) German is consistently harder than English under zero-shot prompting, though structured prompting largely closes this accuracy gap; (ii) larger, more recent, and reasoning-oriented models exhibit stronger identification performance, but are not necessarily more robust to adversarial cues; (iii) incorporating CoT, self-consistency, and few-shot prompting improves accuracy for most models, yet has no consistent effect on ND_M; and finally, (iv) overall, ND_M shows no reliable correlation with scale, reasoning capability, or prompting sophistication, making adversarial vulnerability difficult to predict. Taken together, these findings indicate that while idiom identification benefits from scale, reasoning, and advanced prompting, susceptibility to adversarial context remains largely unpredictable and decoupled from accuracy.

To assess the impact of adversarial cues, we perform an ablation study in which the contexts are rewritten to be neutral, removing adversarial signals (see Appendix G). Overall, longer contexts already challenge idiom identification, but our hard variants amplify this effect by more than 50%.

Additionally, we conduct a qualitative explainability analysis of how LLMs' internal representations respond to adversarial cues. Across 10 PIEs, correctly resolved variants show attention patterns aligned with human-expected literal interpretations, while misleading variants exhibit increased attention from the PIE tokens to adversarial cues. Details and an example appear in Appendix H.

5.1 Human Performance Baseline

To assess whether humans face similar challenges when determining PIEs' usage under the extended-context conditions used in our dataset, we conducted a human evaluation on 178 hard English variants, one per original sentence. Two English-proficient annotators independently labeled each sentence as figurative or literal (with an ambiguous option available), with respect to the usage of the PIE within it. Against the ground truth, the two annotators achieved accuracies of 90.4% and 87.1%, respectively. Inter-annotator agreement reached an accuracy of 81.5%. Both annotators found literal uses harder to classify than figurative ones, consistent with the difficulty our models exhibit, and tended to resolve uncertainty by marking items as ambiguous rather than literal. These results demonstrate that, while adversarial contexts introduce some difficulty, human annotators remain largely robust to misleading cues (90.4% and 87.1% accuracies), correctly classifying the vast majority of cases where models frequently fail.

6 Conclusions

We introduce **ID10M-JAM**, a new adversarial benchmark for idiom identification that systematically probes LLMs under deliberately confusing contexts that remain unambiguous to human readers. By enriching original ID10M sentences with contextual cues favoring the opposite interpretation of a PIE, we stress-test models' contextual and semantic reasoning. This design reveals failure cases that are trivial for humans to avoid but remain hidden under standard evaluation setups. Human evaluation confirms this gap: expert annotators maintain high accuracy ($\sim$90%) on our adversarial variants, demonstrating that the contextual manipulations remain resolvable for humans while severely degrading model performance. Our experiments across English and German reveal that even strong state-of-the-art models, including reasoning-oriented models, remain vulnerable to misleading contextual signals. While sophisticated prompting strategies substantially improve overall identification performance, they do not mitigate susceptibility to adversarial context. Notably, model scale and explicit reasoning mechanisms correlate with higher identification accuracy, yet do not reliably translate into greater robustness, indicating that correct identification and adversarial resilience cannot be seen as coupled. This finding suggests that current models remain overly sensitive to surface-level contextual cues and struggle to maintain stable semantic interpretations under controlled perturbations. ID10M-JAM, therefore, complements existing idiom benchmarks by shifting the focus from performance under clean conditions to robustness under semantic pressure and could be used alongside the standard dataset for idiom identification evaluation. Our methodology offers a general framework for stress-testing contextual understanding in settings where humans remain robust, but models fail. LLMs are often claimed to match humans in language understanding, and our approach can be incorporated into more tasks and settings to challenge them with adversarial edge cases.

Limitations

While our study advances the testing of LLMs for idioms and introduces an adversarial dataset, it also has several limitations. The main limitation is that we included only the English and German splits—leaving Italian and Spanish from the ID10M dataset, and potentially more datasets, for future research. This limits the multilingual insights we can draw from this work. Our source dataset, ID10M, contained only 200 samples per language, so the variability of our produced dataset is limited. Our reliance on ID10M also constrains idiom diversity and may propagate dataset-specific biases. While ID10M provides well-established baselines that allowed us to focus on the adversarial methodology, future work should extend this approach to broader idiom resources. Moreover, during annotation, some original sentences and some generated variants were discarded, further restricting the dataset's scope. Due to budget constraints, we restricted our experiments to a representative subset of configurations: we used three random seeds for the more cost-efficient models and a single one for the more expensive ones. We also selected a limited set of models to represent broader model families, rather than exhaustively evaluating all available options. Because some models are proprietary, reproducibility depends on the availability and stability provided by their providers. A potential concern with using LLMs on existing, publicly available tasks is that the models may have been trained on that data. To partially address this, we evaluate several LLMs, aiming to unveil trends rather than a model's specific performance.

Additionally, our work focuses on diagnosing robustness issues rather than proposing mitigation strategies. Developing concrete methods to improve model performance under adversarial contexts, particularly model-level or training-time interventions beyond prompt engineering, remains an important direction for future work.

These limitations point to several promising directions for future work, including broader evaluations across more datasets, more languages, and more settings.

Ethics Statement

We use publicly available datasets and models in accordance with their intended use, as detailed by the respective publishers and under their licenses; details are listed in Table 7 in Appendix I and in Table 8 in Appendix J. No personally identifiable information or offensive data are processed, and annotators ensure that none exists in the released dataset. Our work is intended for research purposes only, and we see no potential risks. ID10M-JAM, as a derivative of the ID10M dataset, adheres to the same licensing and terms of use of the original publication (Tedeschi et al., 2022) and may be used under the CC BY-NC-SA 4.0 License.

Acknowledgments

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

We would also like to thank Tamar Bar, Ariel Wyrobnik, and Noa Wyrobnik for their work as data annotators in this project and their constant support and availability. We also wish to thank Lihu Zur for his support in revising the data and polishing the paper.

References

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

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*, Nice, France.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational*

Linguistics, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amir Cohen, Hila Gonen, Ori Shapira, Ran Levy, and Yoav Goldberg. 2022. McPhraSy: Multi-context phrase similarity and clustering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3538–3550, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

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

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

Thomas C. Cooper. 1998. Teaching idioms. *Foreign Language Annals*, 31(2):255–266.

Thomas C. Cooper. 1999. Processing of idioms by l2 learners of English. *TESOL Quarterly*, 33(2):233–262.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

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

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

Andrew Gomes. 2025. Anatomy of an idiom: Tracing non-compositionality in language models. *Preprint*, arXiv:2511.16467.

Hila Gonen, Terra Blevins, Alisa Liu, Luke Zettlemoyer, and Noah A. Smith. 2025. Does liking yellow imply driving a school bus? semantic leakage in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 785–798, Albuquerque, New Mexico. Association for Computational Linguistics.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defences and robustness in NLP. *Preprint*, arXiv:2203.06414.

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.

Najet Hadj Mohamed, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar Keskes, and Lamia Hadrich-Belguith. 2024. Lexicons gain the upper hand in Arabic MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 88–97, Torino, Italia. ELRA and ICCL.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

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

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

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Horbach, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner, and Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 836–841, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Ray S. Jackendoff. 1997. *The Architecture of the Language Faculty*, volume 28 of *Linguistic Inquiry Monographs*. MIT Press, Cambridge, MA; London, England.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. *Preprint*, arXiv:1907.11932.

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 21689–21710, Suzhou, China. Association for Computational Linguistics.

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

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile text embeddings distilled from large language models. *Preprint*, arXiv:2403.20327.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. *Preprint*, arXiv:2009.07502.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *Preprint*, arXiv:2310.07849.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of

prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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

Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633.

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

Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An, Xiaonan Wang, Gyuri Choi, and Hansaem Kim. 2025. FLUID QA: A multilingual benchmark for figurative language usage in dialogue across English, Chinese, and Korean. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30268–30282, Suzhou, China. Association for Computational Linguistics.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Howard R. Pollio, John M. Barlow, Howard J. Fine, and Marilyn R. Pollio. 1977. *Psychology and the Poetics of Growth: Figurative Language in Psychology, Psychotherapy, and Education*. Lawrence Erlbaum, Hillsdale, NJ.

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

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden. Association for Computational Linguistics.

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

Prateek Saxena and Soma Paul. 2020. EPIE dataset: A corpus for possible idiomatic expressions. *Preprint*, arXiv:2006.09479.

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

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Varshini Subhash, Anna Bialas, Weiwei Pan, and Finale Doshi-Velez. 2023. Why do universal adversarial attacks work on large language models?: Geometry might be the answer. *Preprint*, arXiv:2309.00254.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Joshua Tanner and Jacob Hoffman. 2023. MWE as WSD: Solving multiword expression identification with word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 181–193, Singapore. Association for Computational Linguistics.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

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

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

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

Su Nam Kim Timothy Baldwin. 2010. *Handbook of Natural Language Processing*, chapter 2:267-292.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2021. Universal adversarial triggers for attacking and analyzing NLP. *Preprint*, arXiv:1908.07125.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Uriel Weinreich. 1969. Problems in the analysis of idioms. In Jaan Puhvel, editor, *Substance and Structure of Language*, pages 23–81. University of California Press, Berkeley and Los Angeles.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Parameter	Other Models
Epochs	20
Batch size	32
Learning rate	2e-5

Table 4: Encoders fine-tuning hyper-parameters.

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

Xingyi Zhao, Lu Zhang, Depeng Xu, and Shuhan Yuan. 2022. Generating textual adversaries with minimal perturbation. *Preprint*, arXiv:2211.06571.

A ID10M-JAM Statistics

The automatic generation and validation process of ID10M-JAM, including model selection, prompt optimization, and experimentation, required approximately 3,000 API calls and cost approximately 20 USD.

B Annotators Information

For the English split, a female high school student proficient in English volunteered for the annotation task and was recruited through one of the paper's authors. One of the paper's authors subsequently reviewed all examples and corrected remaining issues. For the German part, two native German speakers volunteered: one male and one female, both undergraduate students recruited via a call for volunteers at Reichman University. The two annotators split the workload rather than double-annotating the same instances, making the annotation setups comparable between English and German. In both languages, difficult cases were resolved in discussions with the paper's authors.

C Prompt

We use the same prompt as in Hashiloni et al. (2025) for the idiom identification task, see Figure 4.

D Encoder Fine-tuning

The listed hyperparameters are selected based on preliminary experimentation and are presented in Table 4; all remaining settings follow the default configuration of the `Trainer`[8] class in the

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

Model input

System prompt

You are a professional linguist specializing in figurative language and your task is to analyse sentences that may contain an idiom, also known as an idiomatic expression. This is a definition of idiom: 'A phrase, expression, or group of words that has a meaning different from the individual meanings of the words themselves, and employed to convey ideas in a non-literal or metaphorical manner'. Mark idioms only when their usage in the context is idiomatic/figurative and let literal meanings remain unmarked. You are given one sentence in {language}, you are an expert of this language.
If detected, write the idioms exactly as they are in the sentence, without any changes. Only answer in JSON.

Human: Sentence: They've pissed off and left us in the lurch!
AI: idioms: [pissed off]
...

User prompt

Sentence: *Sentence*

Figure 4: Idiom identification prompt, as in Hashiloni et al. (2025).

`transformers` library. The exact code and package versions required are published in the project's repository.

We conduct our fine-tuning experiments on an NVIDIA GeForce RTX 3090 with 24 GB of memory. Overall, all runs and tests, including model selection, took approximately 150 hours.

E Label Studio Example

We provide an example of the Label Studio annotation platform we use for the annotations in Figure 5.

F Full results

In Table 5, we present additional results from our experiments. Specifically, we provide information about the success measured for the models, based on division to literal (L) and idiomatic (I) usages. We display $S_L(M), S_I(M), ND_L(M), ND_I(M)$ accordingly, calculated as explained in Section 4.1. We see a striking asymmetry: $ND_L(M)$ substantially exceeds $ND_I(M)$ across virtually all models and languages. Models that correctly identify a PIE as *literal* in the original sentences are far more likely to flip to an idiomatic prediction under adversarial context than vice versa. This indicates a strong model prior toward idiomaticity—when misleading figurative context is introduced, models tend to over-predict idiomatic usage, making correctly-classified literal PIEs disproportionately vulnerable to adversarial manipulation. Note that due to the low proportion of literal samples (20/178=11.24% in English and 8/137=5.84% in German), those results must be taken with a grain of salt.

In Table 6, we present, for each model and prompting configuration, the distribution of confusing variants induced by each original sentence. There are three potential categories for each sentence s that the model got correctly: (1) AC (All-Confused) are sentences, whose all variants resulted in confusion; (2) NC (None-Confused) are those with no associated confusing variant; (3) MX (Mixed) had some, but not all, of their variants confusing the model. We show an $ACvs.MX$ plot with `Zero-shot` settings across different models and languages in Figure 6, without the encoders and with no separation between the two prompting configurations. English models cluster upper-left—adversarial context is uniformly effective, fooled by either all variants or none. German models spread along the MX axis, indicating greater sensitivity to the specific phrasing of the adversarial cue. Larger and reasoning models show lower AC counts yet non-trivial MX, confirming that vulnerability persists regardless of model strength.

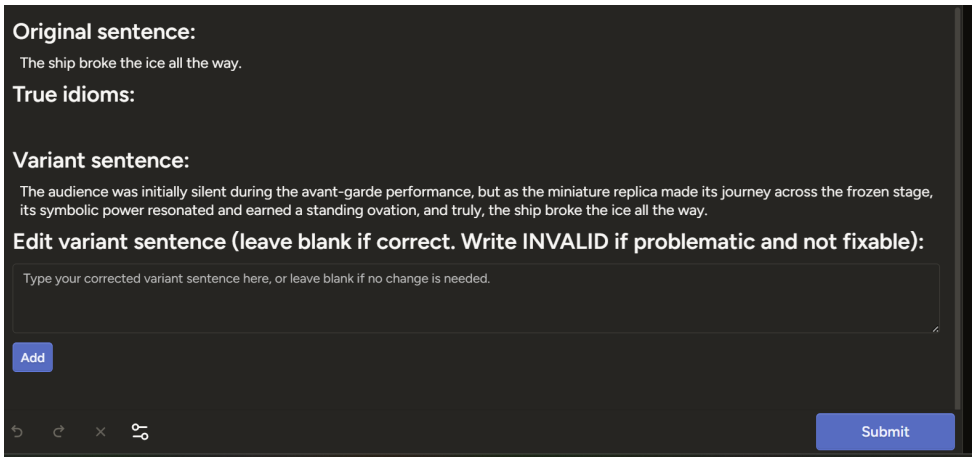


Figure 5: Label Studio annotation example.

		English					German				
Setting	**Model**	S_M	$S_L(M)$	$ND_L(M)(\%)\downarrow$	$S_I(M)$	$ND_I(M)\downarrow$	S_M	$S_L(M)(\%)$	$ND_L(M)(\%)\downarrow$	$S_I(M)$	$ND_I(M)(\%)\downarrow$
Zero-shot	GPT-4o mini	471.00±7.94	45.00±0.00	43.67±1.27	426.00±7.94	1.10±1.04	257.00±4.58	21.00±0.00	28.60±0.00	236.00±4.58	3.80±0.46
	Qwen2.5-72B	452.00±1.73	9.00±0.00	66.70±0.00	443.00±1.73	0.40±0.36	361.00±1.73	0.00±0.00	0.00±0.00	361.00±1.73	9.83±5.12
	Llama 4 Scout	481.00±6.93	44.00±3.46	23.73±4.90	437.00±3.46	4.57±0.78	292.00±6.24	18.00±0.00	25.93±8.46	274.00±6.24	7.77±1.01
	GPT-4o	483.00	45.00	57.80	438.00	1.10	333.00	21.00	19.00	312.00	11.50
	Claude 4 Sonnet	486.00	42.00	76.20	444.00	0.00	369.00	12.00	83.30	357.00	0.30
	Gemini 2.5 Flash	492.00±5.20	41.00±1.73	98.33±1.45	451.00±3.46	0.97±0.46	295.00±1.73	17.00±1.73	79.97±5.77	278.00±1.73	5.23±0.81
	Gemini 2.5 Pro	501.00	60.00	45.00	441.00	2.70	381.00	18.00	38.90	363.00	1.40
Few-shot +SC+CoTBest	GPT-4o mini	483.00±13.75	48.00±0.00	47.20±12.21	435.00±13.75	1.70±0.50	321.00±9.00	23.00±1.73	50.80±1.39	298.00±10.54	8.23±1.40
	Qwen2.5-72B	458.00±1.73	9.00±0.00	29.60±6.41	449.00±1.73	0.33±0.12	380.00±1.73	0.00±0.00	0.00±0.00	380.00±1.73	6.77±1.06
	Llama 4 Scout	487.00±3.46	45.00±5.20	21.90±5.50	442.00±4.58	4.83±0.61	308.00±13.86	18.00±0.00	42.57±8.50	290.00±13.86	6.60±1.68
	GPT-4o	495.00	51.00	51.00	444.00	0.90	390.00	21.00	57.10	369.00	6.00
	Claude 4 Sonnet	498.00	57.00	26.30	441.00	6.10	393.00	21.00	42.90	372.00	5.10
	Gemini 2.5 Flash	484.00±4.58	45.00±0.00	76.30±7.77	439.00±4.58	1.27±0.29	357.00±3.00	14.00±4.58	96.30±6.41	343.00±4.58	4.07±0.78
	Gemini 2.5 Pro	504.00	57.00	52.60	447.00	4.70	390.00	21.00	47.60	369.00	2.40
Reasoning LLMs	DeepSeek-R1	462.00	15.00	86.70	447.00	1.10	378.00	21.00	52.40	357.00	2.00
	o3-mini	477.00	51.00	33.30	426.00	2.10	324.00	24.00	41.70	300.00	6.30
Encoders	mBERT	298.8±6.91	50.40±2.51	34.42±7.51	284.4±8.05	21.80±5.00	206.40±20.28	6.00±0.00	33.30±0.00	200.40±20.28	19.38±1.89
	BERT	357.00±6.00	54.00±0.00	18.54±1.85	303.00±6.00	10.72±0.56	-	-	-	-	-
	GBERT	-	-	-	-	-	313.20±3.42	10.80±1.64	9.42±1.53	302.40±3.29	9.18±1.37

Table 5: More results on the ID10M-JAM dataset. Success count, based on division into literal (L) and idiomatic (I) usages. $S_L(M), S_I(M), ND_L(M), ND_I(M)$ accordingly, calculated as explained in Section 4.1. SC = Self-Consistency, CoT = Chain-of-Thought. When applicable, standard deviation is reported after the ± sign.

G Neutral Context Ablation

To assess the efficacy of adversarial cues, we conduct an ablation study in which they are replaced with neutral contextual prefixes. For each variant v of a sentence s that confused GPT-4o mini under a fixed seed (58 variants in total), we generate a corresponding neutral variant v'. When prompting Gemini to produce v', we keep the instruction as close as possible to the original, while removing adversarial constraints and ensuring that the added context remains semantically coherent with s.

We then evaluate GPT-4o mini on these neutral variants. Despite the absence of adversarial cues, 25 out of 58 variants (43.1%) still lead to confusion. This suggests that increased text length alone can substantially affect idiom identification, even without explicitly adversarial signals. Consequently, these findings raise questions about the robustness of current datasets and indicate that longer, more context-rich inputs may be required for reliable evaluation of this task. At the same time, the substantial performance gap confirms that adversarial cues remain highly disruptive, significantly increasing the confusion rate (ND) over the neutral variants.

H Explainability Analysis

We conduct a qualitative explainability analysis to investigate how LLMs encode PIEs and how adversarial contextual cues influence their internal representations and attention patterns. Following prior probing work (Conneau et al., 2018; Hewitt and Manning, 2019), we analyze Transformer models layer by layer. In addition, similarly to Gomes (2025) we inspect individual attention heads to identify patterns that may be associated with sensitivity to non-compositional language.

We focus on idioms that admit both plausible

		English			German		
Setting	**Model**	AC	NC	MX	A	N	M
Zero-shot	GPT-4o mini	12.00±0.00	153.00±0.00	13.00±0.00	24.33±1.15	90.67±1.15	22.00±2.00
	Qwen2.5-72B	11.00±0.00	156.67±1.53	10.33±1.53	8.33±4.93	106.33±3.21	22.33±4.16
	Llama 4 Scout	5.67±0.58	149.67±2.52	22.67±2.52	14.33±1.53	89.33±1.53	33.33±2.31
	GPT-4o	14.00	151.00	13.00	21.00	91.00	25.00
	Claude 4 Sonnet	16.00	154.00	8.00	6.00	123.00	8.00
	Gemini 2.5 Flash	22.67±1.15	152.00±1.00	3.33±0.58	15.00±1.73	90.67±1.53	31.33±1.15
	Gemini 2.5 Pro	7.00	149.00	22.00	5.00	118.00	14.00
Few-shot	GPT-4o mini	10.67±1.15	149.33±1.53	18.00±1.00	20.00±2.65	87.67±2.31	29.33±2.08
+SC+CoTBest	Qwen2.5-72B	8.33±1.53	156.33±3.21	13.33±2.08	5.00±1.00	105.33±3.06	26.67±2.52
	Llama 4 Scout	9.00±2.00	156.67±3.79	12.33±5.51	8.33±0.58	106.00±2.65	22.67±2.52
	GPT-4o	10.00	151.00	17.00	5.00	113.00	19.00
	Claude 4 Sonnet	9.00	144.00	25.00	4.00	107.00	26.00
	Gemini 2.5 Flash	15.33±0.58	148.67±4.51	14.00±4.58	9.33±1.53	107.33±3.06	20.33±2.08
	Gemini 2.5 Pro	9.00	141.00	28.00	4.00	117.00	16.00
Reasoning	DeepSeek-R1	17.00	149.00	12.00	1.00	121.00	15.00
LLMs	o3-mini	11.00	153.00	14.00	21.00	95.00	21.00
Encoders	mBERT	71.60±11.19	227.20±11.50	0.00±0.00	40.60±2.88	165.8±19.56	0.00±0.00
	BERT	42.40±1.52	314.60±6.31	0.00±0.00	-	-	-
	GBERT	-	-	-	28.80±4.49	284.40±2.41	0.00±0.00

Table 6: Distribution of confusing variants count per sentence. (1) AC (All-Confused) are sentences, that all their variants resulted in a confusion; (2) NC (None-Confused) are those with no associated confusing variant; (3) MX (Mixed) had some, but not all, of their variants confusing the model. SC = Self-Consistency, CoT = Chain-of-Thought. When applicable, standard deviation is reported after the ± sign.

literal and idiomatic interpretations and select 10 paired sentence variants with comparable surface forms. Specifically, we analyze variants that are correctly resolved by all evaluated models (COR), and variants that mislead at least a subset of models (MIS). All sentences are passed through Llama-3.2-3B, and attention distributions are manually examined across layers and heads.

A consistent qualitative pattern emerges. In COR variants, tokens corresponding to the PIE (e.g., *break*, *ice*) primarily attend to tokens that align with a human-expected literal interpretation. By contrast, in MIS variants, these tokens disproportionately attend to adversarial contextual cues, which may lead to misplaced salience and, downstream, to incorrect idiomaticity judgments. Figure 7 illustrates a representative simplified example of this behavior.

I Model Checkpoints

In Table 7, we present the checkpoints (or snapshots) used in this work and the models' sizes and licenses.

J Artifacts

We detail artifacts we use and their respective usage and licenses in Table 8. For the exact versions of and more details, see the *requirements.txt* file in the project's repository.

We implement our prompting framework with LangChain and agno because they provide a modular and reproducible framework for LLM evaluation, particularly when working with multiple providers, as it wraps their APIs in a unified layer.

K AI assistants

We used AI assistants (e.g., ChatGPT) to support code formatting, phrasing suggestions, and LaTeX styling during writing. All outputs were reviewed and edited by the authors. No content was directly generated or used without human verification.

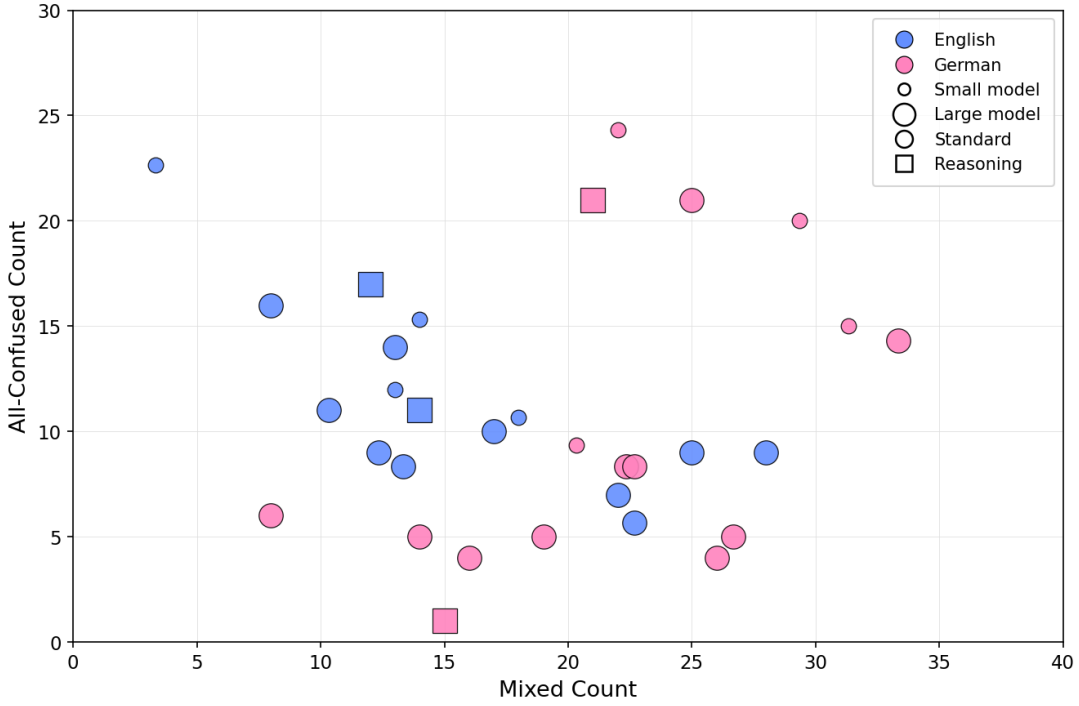


Figure 6: Visualization of AC, MX (All-Confused and Mixed) categories across different languages and models. We do not distinguish between the two prompting configurations for this purpose. Encoders are excluded from this plot.

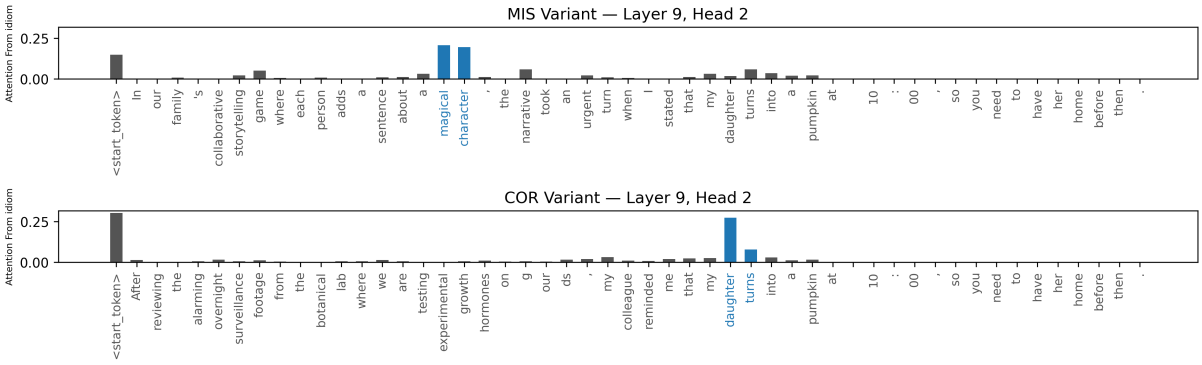


Figure 7: Attention score: the query is the idiom tokens and the keys are the context tokens. Upper is a `MIS` variant, and bottom is a `COR` variant, rooted from the same original sentence.

Model	Checkpoint	License	# Param
GPT-4o mini[a]	gpt-4o-mini-2024-07-18	Proprietary	>200B∗
GPT-4o[b]	gpt-4o-2024-08-06	Proprietary	8B∗
o3-mini[c]	o3-mini-2025-01-31	Proprietary	200B∗
Claude 4 Sonnet[d]	claude-sonnet-4-20250514	Proprietary	150-250B∗
Gemini 2.5 Flash[e]	gemini-2.5-flash	Proprietary	N/A
Gemini 2.5 Pro[f]	gemini-2.5-pro	Proprietary	N/A
Llama 4 Scout[g]	Llama-4-Scout-17B-16E-Instruct	Llama 4 Community License Agreement	17B active 109B overall
Qwen2.5-72B [h]	Qwen2.5-72B-Instruct-Turbo	Qwen LICENSE	72B
DeepSeek-R1[i]	deepseek-ai/DeepSeek-R1-0528	MIT License	671B
mBERT[j]	google-bert/bert-base-multilingual-cased	Apache-2.0	110M
BERT[k]	google/bert-bert-base-uncased	Apache-2.0	110M
GBERT[l]	deepset/gbert-base	MIT License	110M
Llama-3.2-3B[m]	meta-llama/Llama-3.2-3B	Llama 3.2 Community License Agreement	3B

[a] https://platform.openai.com/docs/models/gpt-4o-mini
[b] https://platform.openai.com/docs/models/gpt-4o
[c] https://platform.openai.com/docs/models/o3-mini
[d] https://www.anthropic.com/news/claude-4
[e] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash
[f] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro
[g] https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E
[h] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[i] https://huggingface.co/deepseek-ai/DeepSeek-R1-0528
[j] https://huggingface.co/google-bert/bert-base-multilingual-cased
[k] https://huggingface.co/google-bert/bert-base-uncased
[l] https://huggingface.co/deepset/gbert-base
[m] https://huggingface.co/meta-llama/Llama-3.2-3B

Table 7: Checkpoints used during experiments and their License, and their number of parameters. ∗ = non-official estimation, as this information is not public. N/A means not disclosed, and an estimation can't be found.

Artifact	Type	License	Usage
LangChain[a]	Framework	MIT License	Prompting
Agno[b]	Framework	Apache-2.0	Prompting
Together AI[c]	Provider	Proprietary	API access
Label Studio AI[d]	Platform	Apache-2.0	Annotations
ID10M[e]	Dataset	CC BY-NC-SA 4.0	Source data

[a] https://www.langchain.com/
[b] https://www.agno.com/
[c] https://www.together.ai/
[d] https://labelstud.io/
[e] https://github.com/Babelscape/ID10M

Table 8: Packages, dataset and artifacts used during experiments, along with their license and usage explanation.