Multilingual Idioms in Sentences and Conversations Across High-, Medium-, and Low-Resource Languages

Anonymous ACL submission

Abstract

Idiomatic expressions pose a major challenge for multilingual NLP because their meanings shift between figurative and literal usage, often requiring context for accurate interpretation. Prior work has focused on high-resource languages typically evaluates isolated idiom-meaning questions, overlooking realistic discourse. We introduce MIDI, a multilingual idiom dataset spanning 3 high-, 3 medium-, and 12 low-resource languages, curated by native speakers. Unlike previous datasets, MIDI provides idioms embedded in both sentence-level and conversational contexts, capturing both literal and figurative readings. Benchmarking state-of-the-art models shows that idiom comprehension degrades in low-resource languages and that, in all resource tiers, literal interpretations are substantially harder than figurative ones. Conversational context improves performance but does not eliminate these disparities. Through controlled tests and interventions on hidden representations, we further separate memorization from reasoning, exposing core limitations of current models[1].

1 Introduction

While large language models (LLMs) demonstrate impressive capabilities, processing idioms remains a focus of ongoing research (Zhou and Bhat, 2024; Zhou et al., 2024; Kim et al., 2025) due to their semantic ambiguity between literal and figurative meanings (Baldwin and Kim, 2010). Idioms serve as a unique test bed because, while their meanings can be memorized as patterns, correctly inferring them in context requires integrating nuanced cultural cues and reasoning-based inference (Cacciari and Tabossi, 1988; Dankers et al., 2022; Kovács, 2016).

Recent studies show that state-of-the-art models perform well with high-resource idioms (Mi et al., 2025), but struggle in lower-resource settings where training data is sparse and deep cultural grounding is required. The current study expands this test bed to investigate the complex interplay between memorization and reasoning (Kim et al., 2025). By examining how models handle these non-compositional expressions, we aim to better characterize the hybrid mechanisms that drive idiomatic understanding across diverse linguistic landscapes.

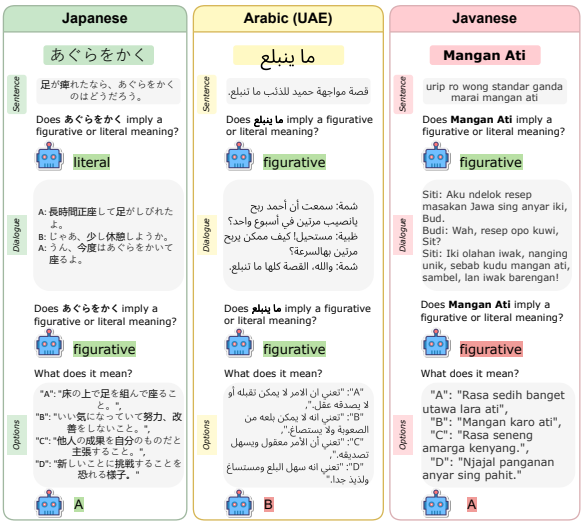

Figure 1: We compile idioms and their sentence- and dialogue-level usages from 18 languages spanning high-, medium-, and low-resource contexts, then evaluate LLMs with multiple-choice and binary inference tasks targeting both figurative vs. literal understanding and biased interpretations.

Existing multilingual benchmarks are limited in their language diversity or the discourse environments where idioms naturally occur (see Table 1 for more details). Besides, current multilingual large language models (mLLMs) have not been systematically tested on their ability to reason over the idioms' semantic ambiguity across diverse linguistic and cultural contexts. This leads us to pose the following research questions that we expect

[1]For reproducibility, we make the dataset and the code openly accessible anonymized repo.

Reference	# of Languages	MCQ	Context Granularity		Low Resource	Multi-dimension annotation
			Sentence	Conversation		
LIDIOMS (Moussallem et al., 2018)	5	✗	✔	✗	✗	✗
MAGPIE (Haagsma et al., 2020)	1	✗	✔	✔	✗	✗
AStitchInLanguageModels (Tayyar Madabushi et al., 2021)	2	✗	✔	✔	✗	✗
ID10MS (Tedeschi et al., 2022)	10	✗	✔	✗	✗	✗
MAPS (Cecilia Liu et al., 2024)	6	✔	✗	✔	✔	✗
Persian-MAPS(Khoshtab et al., 2025)	1	✔	✗	✔	✔	✗
DICE (Mi et al., 2025)	1	✗	✔	✗	✗	✗
MIDAS (Kim et al., 2025)	6	✔	✔	✗	✗	✗
Ours	18	✔	✔	✔	✔	✔

Table 1: **MIDI vs. prior idiom benchmarks** across languages, context, coverage, and annotations.

to answer with a view of diverse linguistic landscapes: *(1) Do models interpret idioms by memorizing their meanings, or by reasoning from context to disambiguate usage? (2) Do they generalize across languages, or does idiom disambiguation remain language-specific? (3) To what extent does context—sentential or conversational—help models separate figurative and non-figurative meaning?*

To address these questions, we curate a multilingual idiom dataset **MIDI** spanning **18 languages and dialects**. Following the taxonomy established by Joshi et al. (2020), we categorized these languages into three tiers based on their digital availability and institutional support: high-resource *(Chinese, Japanese, Russian)*, medium-resource (*Indonesian, Vietnamese, and the UAE dialect of Arabic*) languages demonstrating moderate digital presence and are identified as "Rising Stars" in global NLP benchmarks (Costa-Jussà et al., 2022), and 12 low-resource languages and regional dialects *(Javanese, Persian, Kannada, Telugu, Tamil, Minangkabau, Sundanese, Kazakh, Yoruba, Arabic [Syrian, Egyptian, and Moroccan dialects])*, selected due to their severe underrepresentation in standard training sets and their reliance on nuanced cultural grounding for idiomatic interpretation.

MIDI contains over 100 idioms per language, selected and validated by native speakers and annotated with key psycholinguistic properties, including familiarity, literal plausibility, and idiom decomposability (Libben and Titone, 2008). Importantly, the idioms in the dataset are embedded in two realistic discourse settings: (1) sentence-level contexts, each crafted in both literal and figurative senses; and (2) short dialogues, where pragmatic cues, speaker intentions, and conversational flow influence interpretation. Each instance is paired with manually verified multiple-choice comprehension questions for controlled evaluation.

To further investigate memorized idiomatic knowledge vs. genuine reasoning, we provide parallel evaluation splits of "memorization" and "contextual reasoning," inspired by recent methodology that separates definitional recall from usage-based inference (Kim et al., 2025). By probing models on figurative/literal classification both with and without prior access to idiom definitions, we evaluate current systems' abilities to interpret idioms beyond rote recall.

In summary, our contributions are as follows: (i) we introduce **MIDI**, a broad testbed spanning 18 typologically and culturally diverse languages and dialects to study how mLLMs resolve idiomatic ambiguity across high-, mid-, and low-resource settings; (ii) we provide idioms grounded in both literal and figurative sentence contexts as well as multi-turn dialogues; and (iii) we benchmark current LLMs and reveal persistent gaps in low-resource languages and for literal readings; while conversational context often improves accuracy, it does not close these gaps, highlighting persistent limitations in multilingual idiom understanding; (iv) we additionally show that activation steering along memorization and reasoning dimensions yields consistent gains, especially for low-resource languages, with MMLU-Pro (Wang et al., 2024) directions transferring effectively to MIDI, suggesting a shared mechanism behind the memorization–reasoning tradeoff.

2 Related Work

A significant body of literature has introduced various idiom datasets, reflecting a growing interest in the computational modeling of figurative language. MAGPIE (Haagsma et al., 2020) is one of the most popular datasets for English providing 56K instances of potentially idiomatic expressions across 1.7K idioms, focusing primarily on sentence-level annotations without multilingual ex-

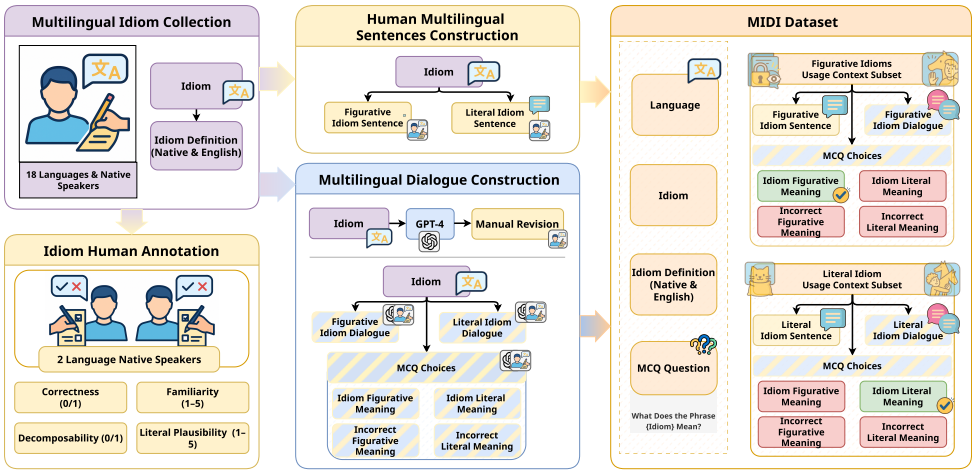

Figure 2: **MIDI construction pipeline.** Native speakers collect and annotate idioms (18 languages) with bilingual definitions, then create figurative/literal sentence contexts and LLM-generated (manually revised) dialogues with MCQ options, producing paired figurative and literal usage subsets.

tension. The more recent DICE dataset (Mi et al., 2025) introduces a contrastive evaluation framework for idiomatic expressions also for English only, examining whether models can correctly interpret idioms in context.

Lai et al. (2023) propose multilingual multi-figurative language detection using a unified prompt-based framework, demonstrating generalization across languages, figures of speech, and zero-shot settings. Khoshtab et al. (2025) compare how different large language models interpret idioms and similes across multiple languages — especially including new Persian datasets— and how various prompting strategies affect their performance. Beyond idioms, proverb interpretation has served as a lens for examining cultural knowledge and reasoning in multilingual language models. MAPS (Cecilia Liu et al., 2024) introduces a multicultural dataset of proverbs and sayings across six languages in conversational contexts, providing analyses of memorization versus reasoning and highlighting cultural gaps in mLLMs.

Prior work also explores the binary classification of literal versus figurative usage (Tedeschi et al., 2022; De Luca Fornaciari et al., 2024), establishing the importance of context for idiom interpretation with evaluations restricted to English or a small set of languages.

Table 1 summarizes the comparison of the existing datasets (Moussallem et al., 2018; Haagsma et al., 2020; Tayyar Madabushi et al., 2021; Tedeschi et al., 2022; Kim et al., 2025) with **MIDI** along 5 dimensions: (1) *Number of languages:* **MIDI** covers the largest and most diverse set of languages to date, (2) *MCQ availability*: only one of 5 works provides multiple-choice questions for evaluating idioms, (3) *Context Granularity:* This dimension assesses the contextual granularity of the dataset, specifically whether it provides sentence-level or conversation-level information, (4) *Low Resource:* this dimension relates to whether the languages that are considered low-resource, and (5) *Multi-dimensional idiom Annotation:* this checks if additional human annotations (beyond simple definitions) related to the idiom are available; in **MIDI** we include 3 additional psycholinguistic properties of an idiom.

3 Dataset Construction

The process starts with multilingual idiom collection across 18 languages and dialects, carried out by native speakers, experts in linguistics, from different age groups, both male and female. The languages included in the dataset and their statistics are shown in Table 2. For each language, annotators collected idioms that are commonly used and familiar to native speakers.

MIDI contains 2,278 idioms and 8,460 usage contexts, with each idiom appearing in both as a sentence and a dialogue. On average, 82% of the idioms have a literal counterpart, though this varies across languages. Some languages show a high overlap between figurative and literal meanings, while others do not. In particular, Vietnamese idioms in the dataset are exclusively non-literal, and languages such as Yoruba and Minangkabau contain relatively few idioms with literal interpretations.

Language	#Idioms	#Literal (%)	#Contexts
High-Resource			
Chinese	300	100	1200
Japanese	115	95	448
Russian	213	98	842
Mid-Resource			
Arabic (UAE)	100	98	396
Indonesian	108	91	412
Vietnamese	100	0	200
Low-Resource			
Arabic (Egypt)	100	84	368
Arabic (Morocco)	99	94	384
Arabic (Syria)	100	100	400
Persian	102	80	368
Javanese	104	92	400
Kannada	198	100	792
Kazakh	100	100	400
Minangkabau	100	49	298
Sundanese	100	72	344
Tamil	99	96	388
Telugu	139	99	552
Yoruba	101	33	268
Total / Avg.	**2,278**	**82**	**8,460**

Table 2: Languages in **MIDI** by resource tier (High / Mid / Low). #Literal is the share of figurative idioms with a literal counterpart; #Contexts counts sentence and dialogue instances (figurative+literal).

Dimension Annotation Each idiom is validated by at least two native speakers and annotated following Libben and Titone (2008): decomposability (binary), familiarity (1–5), and literal plausibility (1–5).

Figurative and Literal Example Sentences the collected idioms are supported by two example sentences: one illustrating the idiom figurative usage and another demonstrating its literal, word-by-word interpretation, context. The sentences are manually authored (and also manually curated) and do not originate from existing sources, ensuring no data leakage into the training process.

Multilingual Dialogue Construction For each idiom, we asked GPT-5 to generate two types of dialogues: a dialogue with the idiom in figurative meaning and another with its literal meaning. The prompts for each dialogue type are presented in Appendix B. In addition to the dialogues, we ask to generate multiple-choice questions (MCQs) to further probe idiom understanding. The answer options are the correct figurative meaning, the correct literal meaning, and corresponding incorrect distractors. All generated dialogues and MCQs undergo a manual revision step by the same annotators from previous step to ensure linguistic quality and semantic correctness. However, for some low-resource languages such as Minangkabau and Yoruba, the LLM-generated text was almost unusable, requiring annotators to rewrite nearly 100% of the content from scratch.

Final dataset The collected data is organized into the multilingual idioms dataset **MIDI**. Each instance is explicitly associated with its language, idiom, and bilingual (native language and English) definition. Overall, **MIDI** consists of two complementary subsets: a figurative idiom usage context subset and a literal idiom usage context subset. Each subset contains the corresponding idiom sentence, idiom dialogue, and an MCQ asking *"What does the phrase [IDIOM] mean?"*, along with answer choices. In the figurative subset, the figurative meaning is marked as correct, while in the literal subset, the literal meaning is marked as correct. Instances of the **MIDI** dataset are given in Table 16.

4 Experiment

Using **MIDI**, we benchmark multilingual LLMs to characterize idiom comprehension across (i) language resource levels, (ii) sentence vs. conversational contexts, and (iii) figurative vs. literal usage.

4.1 Experimental Setup

We evaluate both proprietary and open-source LLMs representing the current state of the art in multilingual NLP. The proprietary models include GPT-5.2 (OpenAI, 2025) and Gemini 2.5 Pro (Comanici et al., 2025), which represent frontier capabilities in multilingual reasoning and understanding. The open-source models include DeepSeek-R1-Distill-Llama (70B) (DeepSeek-AI et al., 2025), Gemma-3 (27B Instruct) (Team et al., 2025), Llama-3.1 (8B Instruct), Llama-3.3 (70B Instruct) (Grattafiori et al., 2024), Mixtral (8×7B Instruct) (Jiang et al., 2024), Qwen-3 (4B Instruct; 30B-MoE Instruct) (Yang et al., 2025). This set spans a range of sizes and architectures, enabling analysis of how scale and design choices relate to idiom understanding.

For evaluation we cast idiom comprehension (*reasoning + memorization*) as a four-way multiple-choice question (MCQ) task. Given an idiom embedded in context (sentence-level or multi-turn dialogue), the model selects the correct interpretation from four options: the figurative meaning, the literal meaning, and two distractors (one figurative, one literal). We report accuracy across idiom usage conditions (figurative or literal), with random baseline performance at 25%. For exact task formulation, refer to the Appendix C.1.

Model	Idiom Usage Context						Idiom Usage Type						Overall
	High		Mid		Low		High		Mid		Low		
	Sent	Conv	Sent	Conv	Sent	Conv	Fig	Lit	Fig	Lit	Fig	Lit	
Random	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Proprietary													
GPT-5.2	65.1	67.1	**90.0**	**91.1**	68.5	77.1	97.0	35.2	97.2	76.8	86.7	58.9	74.4
Gemini 2.5 Pro	67.5	71.2	89.5	90.9	**69.5**	**78.2**	**98.6**	40.2	**99.2**	72.1	**94.8**	52.9	**75.5**
Open-Source													
DeepSeek-R1 (70B)	48.8	54.2	53.6	57.5	34.6	40.5	55.1	47.9	57.5	46.4	42.6	32.6	44.0
Gemma-3 (27B)	67.5	73.2	86.9	90.0	58.8	70.9	96.9	43.8	96.2	71.7	76.4	53.3	71.0
Llama-3.1 (8B)	62.4	68.6	27.6	31.7	39.6	45.0	82.0	49.0	29.3	33.5	43.1	41.6	49.5
Llama-3.3 (70B)	64.6	65.2	65.5	75.2	53.5	62.4	87.5	42.3	77.9	59.4	70.7	45.2	63.6
Mixtral (8×7B)	**69.4**	**76.9**	59.6	75.6	42.6	53.2	79.4	**67.0**	57.9	**80.7**	35.7	**60.1**	57.5
Qwen-3 (4B)	30.6	38.3	31.0	35.1	27.5	28.5	43.3	25.7	36.3	32.8	27.1	28.9	30.8
Qwen-3 (30B-MoE)	40.8	37.4	24.1	23.5	25.6	25.8	43.9	34.3	24.4	23.3	24.1	27.2	31.2
Average	*57.4*	*61.4*	*58.6*	*63.4*	*46.7*	*53.5*	*76.0*	*42.8*	*64.0*	*55.2*	*55.7*	*44.5*	*55.3*

Sent = Sentence *Conv = Conversation* *Fig = Figurative* *Lit = Literal*

Table 3: Model performance (accuracy) on idiom comprehension MCQ across usage contexts and types, stratified by language resource availability (High, Mid, Low). **Bold** indicates best; underline indicates second-best.

All evaluations are conducted in a zero-shot setting to assess models' inherent capabilities without task-specific fine-tuning using lm-evaluation-harness (Gao et al., 2024). We aggregate results across instances and analyze performance along three axes: (i) *context type* (sentence vs. dialogue), (ii) *usage type* (figurative vs. literal), and (iii) *language resource availability level* (high-, medium-, and low-resource). This decomposition isolates the effects of context, semantic ambiguity, and data availability on idiom understanding.

4.2 Idiom Comprehension MCQ

Table 3 presents idiom comprehension results across all evaluation dimensions.

Overall Observation. Proprietary models substantially outperform open-source alternatives: Gemini 2.5 Pro achieves the highest accuracy at 75%, followed closely by GPT-5.2 at 74%, while the best open-source model, Gemma-3 (27B), reaches 71%. At the lower end, Qwen-3 (4B) and Qwen-3 (30B-MoE) perform near random chance both at 31%, indicating that smaller or sparsely-activated architectures struggle with multilingual idiom understanding. Notably, DeepSeek-R1 (70B) underperforms relative to its parameter count, achieving only 44% overall, substantially below the smaller Gemma-3, suggesting that model architecture and training objectives may matter as much as scale. An unexpected pattern emerges in the resource availability level results: medium-resource languages often yield the highest accuracy, GPT-5.2 achieves 90% on mid-resource sentences versus 65% on high-resource, possibly reflecting favorable typological properties or higher annotation quality for these languages.

Model	Memorization			Overall
	High	Mid	Low	
Random	*25.0*	*25.0*	*25.0*	*25.0*
Proprietary				
GPT-5.2	97.8	94.4	84.0	89.3
Gemini 2.5 Pro	**98.8**	**97.7**	**92.5**	**95.0**
Open-Source				
DeepSeek-R1 (70B)	43.5	47.0	34.1	38.4
Gemma-3 (27B)	96.4	86.9	67.3	78.9
Llama-3.1 (8B)	71.4	29.2	43.4	51.4
Llama-3.3 (70B)	38.6	44.9	56.9	51.4
Mixtral (8×7B)	76.8	53.6	43.4	55.4
Qwen-3 (4B)	50.3	30.2	23.7	33.3
Qwen-3 (30B-MoE)	49.7	24.2	22.4	33.4
Average	*69.3*	*56.4*	*52.0*	*58.5*

High, Mid, Low = Resource Level

Table 4: Model accuracy on idiom memorization (figurative meaning identification given no context). **Bold** = best; underline = second-best.

Resource Availability. Performance degrades as language resource availability decreases. Averaging across models, high-resource languages achieve 59% accuracy, medium-resource languages reach 61%, and low-resource languages attain only 50%. This degradation is particularly pronounced for open-source models; for instance, Gemma-3 (27B) achieves 70% on high-resource languages but drops to 65% on low-resource languages. The gap is somewhat narrower for proprietary models, with Gemini 2.5 Pro maintaining relatively robust performance across resource levels (69.4% high, 90.2% mid, 73.9% low), suggesting that larger-scale training may partially mitigate resource scarcity effects. We also observe that medium-resource languages yield particularly strong performance for several models, GPT-5.2 and Gemma-

Model	Reasoning			Overall
	High	Mid	Low	
Random	*25.0*	*25.0*	*25.0*	*25.0*
Proprietary				
GPT-5.2	98.2	98.5	95.6	96.4
Gemini 2.5 Pro	**98.9**	**99.7**	**97.5**	**98.0**
Open-Source				
DeepSeek-R1 (70B)	54.3	61.3	53.9	54.7
Gemma-3 (27B)	97.5	98.9	91.2	93.9
Llama-3.1 (8B)	96.8	54.3	73.7	78.1
Llama-3.3 (70B)	95.9	91.0	77.8	85.2
Mixtral (8×7B)	94.5	91.0	72.2	80.8
Qwen-3 (4B)	44.3	36.8	31.7	36.9
Qwen-3 (30B-MoE)	65.5	35.9	31.3	43.0
Average	*82.9*	*74.1*	*69.4*	*74.1*

High, *Mid*, *Low* = *Resource Level*

Table 5: Model accuracy on idiom reasoning (figurative meaning identification given usage context and English meaning). **Bold** = best; underline = second-best.

3 reach 90% and 87% respectively, on medium-resource sentences vs. 65% and 68% on high-resource sentences, indicating that resource tier is not the only determinant of difficulty.

Context Type. Conversational contexts consistently improve model performance compared to sentence-level contexts. Across all models and resource availability levels, conversation-based evaluation yields an average accuracy of 59% compared to 54% for sentences. This improvement is most pronounced for low-resource languages, where the gap between conversation (54%) and sentence (47%) contexts reaches 7 percentage points. This is consistent with the idea that multi-turn discourse provides additional pragmatic cues including speaker intentions and discourse coherence, that help models disambiguate idiomatic usage even when lexical familiarity is limited.

Figurative vs. Literal Comprehension. Models exhibit substantially higher accuracy on figurative usage interpretation (65% average) compared to literal usage (48% average). This asymmetry is most extreme for high-resource languages, where the gap reaches 33 percentage points (76% figurative vs. 43% literal). The pattern suggests that models more readily identify figurative meanings, likely due to stronger training signal from canonical idiom definitions, but struggle to recognize when contextual cues indicate a compositional literal reading. This highlights a fundamental limitation in current models' ability to flexibly disambiguate idiom usage based on context. Mixtral (8×7B) is a notable exception, achieving the highest literal accuracy across all resource availability levels (67% high, 81% mid, 60% low), indicating stronger sensitivity to literal-signaling context, though at the cost of lower figurative performance in medium and low-resource settings.

These findings show that idiom comprehension remains challenging for current LLMs, particularly in low-resource languages and literal usages. We next analyze the underlying mechanisms: Section 5.1 isolates memorization and contextual reasoning, Section 5.2 examines models' interpretation bias toward figurative readings, and Section 5.3 studies how steering memorization and reasoning affects performance.

5 Analysis

5.1 Memorization and Reasoning

Memorization. The memorization settings evaluates whether a model can recall the figurative meaning of an idiom without any contextual support. In this task, the model is presented with an idiom in isolation and asked *"What is the figurative meaning of this idiom?"*, forcing it to rely solely on its stored knowledge rather than contextual reasoning (see Appendix C for details). Results are shown in Table 4. Proprietary models achieve consistently high performance across all resource levels. This suggests broad memorization of idiomatic expressions across languages. In contrast, open-source models exhibit substantial variation. Larger models such as Gemma-3 and Mixtral show relatively strong memorization in high-resource languages, while smaller models and those evaluated on mid- and low-resource languages perform substantially worse. This pattern mirrors the main results, where limited memorization often coincides with weaker figurative understanding in context.

Reasoning. In contrast to memorization, the reasoning setting examines the model's ability to identify the figurative meaning of idiom when both the usage context and the English meaning are provided. In this task, the model model is no longer required to rely on memorization knowledge alone, but instead must use the given information to reason about the correct interpretation (see Appendix C for prompt details). Results are shown in Table 5. Proprietary models achieve near-perfect accuracy across all resource levels, indicating strong and consistent reasoning ability even in low-resource languages. Meanwhile, open-source models benefit substantially from the availability of context and meaning, with several large models, such as

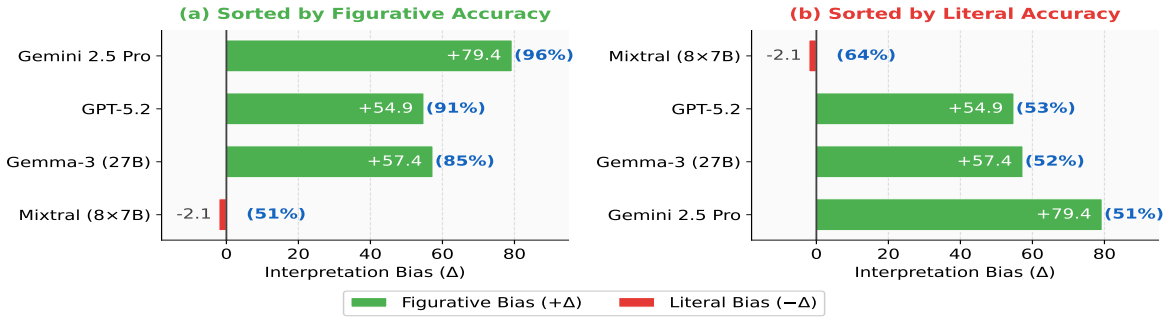

Figure 3: Interpretation bias (Δ = figurative–literal preference, %) vs. idiom comprehension accuracy: models sorted by (a) figurative and (b) literal accuracy (parentheses show accuracies).

Gemma-3 and Mixtral, approaching proprietary performance in high-resource settings, though performance still degrades in mid- and low-resource languages. Compared to memorization, reasoning reduces but does not eliminate the performance gap between proprietary and open-source models, suggesting that context helps compensate for limited idiomatic knowledge but does not fully resolve it. Performance patterns are largely consistent between sentence and conversational contexts, with detailed results reported in Table 9.

Model	Vector Source	Type	High	Mid	Low	Avg.
Llama-3.1 (8B)	MIDI-derived	Mem.	0.25	0.37	0.64	0.44
		Reas.	0.14	0.34	0.60	0.36
	MMLU-Pro	Mem.	0.22	0.32	0.59	0.40
		Reas.	0.20	0.31	0.63	0.38
Qwen-3 (4B)	MIDI-derived	Mem.	0.52	0.34	0.86	0.64
		Reas.	0.32	0.28	0.65	0.40
	MMLU-Pro	Mem.	0.71	0.41	1.26	0.78
		Reas.	0.38	0.36	0.62	0.41

Table 6: Best-layer steering gains (Δ accuracy in percentage points), macro-averaged across languages within each resource tier and subsequently averaged over all tasks.

5.2 Interpretation Bias

To understand the asymmetric figurative-literal performance gap observed in Section 4.2, we analyze models' default interpretation tendencies when presented with ambiguous idiom prompts lacking contextual cues. We query each model with the prompt *"What does the phrase {idiom} mean?"* followed by the figurative and literal meanings as options, without any usage context to guide interpretation (see Appendix C for details).

Figure 3 reveals a strong correlation between interpretation bias and task performance. Models exhibiting strong figurative bias, preferring figurative interpretations in ambiguous contexts, achieve substantially higher accuracy on figurative meaning identification. Gemini 2.5 Pro shows the strongest figurative bias (Δ=+79.4) and achieves the highest figurative accuracy (96%), followed by GPT-5.2 (Δ=+54.9, 91%) and Gemma-3 (Δ=+57.4, 85%). This pattern suggests that models' strong figurative performance stems partly from a default tendency to interpret idioms figuratively, likely reinforced by training data where idioms predominantly appear with figurative meanings.

Conversely, the literal accuracy ranking reveals an inverse pattern: models with weaker figurative bias achieve better literal comprehension. Mixtral, with near-neutral bias (Δ=–2.1), achieves the highest literal accuracy (64%). Among the strong figurative-bias models, GPT-5.2 (53%) and Gemma-3 (52%) outperform Gemini 2.5 Pro (51%) on literal tasks—notably, their literal performance ordering mirrors their bias magnitude, with Gemini's strongest figurative bias (Δ=+79.4) corresponding to the weakest literal comprehension despite its overall superiority on figurative interpretation. This suggests that excessive figurative bias impairs models' ability to recognize literal usage, even for otherwise highly capable models.

Full bias analysis across all models is provided in Appendix F.

5.3 Activation Steering Analysis

Setup. To better understand the interplay between memorization and reasoning in LLMs, we extend our analysis to activation steering, examining how latent representations can generalize and enhance performance in low-resource settings. We perform inference-time *activation steering* by linearly modifying residual-stream activations at a selected layer using activation addition (ActAdd) (Turner et al., 2023). Following the Linear Reasoning Features (LiReFs) framework (Hong

et al., 2025), we derive a steering direction via a difference-in-means estimator. Given prompt sets eliciting reasoning (D_{REAS}) and memorization (D_{MEM}), let $h^{(\ell)}(x)$ denote the residual-stream activation at layer ℓ for the final prompt position. We compute:

$$r^{(\ell)} = \frac{1}{|D_{\text{REAS}}|} \sum_{x \in D_{\text{REAS}}} h^{(\ell)}(x) - \frac{1}{|D_{\text{MEM}}|} \sum_{x \in D_{\text{MEM}}} h^{(\ell)}(x). \quad (1)$$

At inference time, we steer by adding a scaled copy of this direction vector:

$$h^{(\ell)}(x) \leftarrow h^{(\ell)}(x) + \alpha r^{(\ell)}, \quad \text{with } |\alpha| = 0.1. \quad (2)$$

We compare steering directions from two sources: **MIDI-derived vectors**, computed from MIDI prompts, and **MMLU-Pro vectors**, extracted from an independent benchmark (Wang et al., 2024). For each source, we evaluate two steering polarities: *memorization steering* ($\alpha < 0$) and *reasoning steering* ($\alpha > 0$).

Tasks and layer sweep. We evaluate steering across the five MIDI configurations introduced earlier: **Dialogue Standard**, **Dialogue Reasoning**, **Memorization**, **Sentence Standard**, and **Sentence Reasoning**. For computational tractability, we sweep alternate layers beginning at layer 3 ($\ell \in \{3, 5, 7, \dots\}$; through ℓ=31 for Llama-3.1 8B and ℓ=35 for Qwen-3 4B) and report results at the **best-performing layer** selected independently for each combination of language, task, steering type, and vector source.

Main findings. Steering yields **modest but consistent** accuracy gains, with the largest improvements concentrated in **low-resource languages** (Appendix G), consistent with a headroom effect. MMLU-Pro vectors transfer well to idiom understanding, performing comparably to MIDI-derived vectors overall and slightly better for Qwen in aggregate.

Interestingly, memorization steering often yields slightly larger gains than reasoning steering (Table 6). This supports the broader picture from our controlled probes: memorization and reasoning are **not cleanly separable**. For idioms, "memorization" features can act as *sense anchors* (a stored inventory of candidate meanings) that *enable* downstream contextual reasoning and disambiguation rather than competing with it.

Steering is also **layer-sensitive**: the best layer varies across languages and tasks, and suboptimal layers can degrade performance. Notably, the distribution of best layers is similar between MIDI-derived and MMLU-Pro vectors within each model; see Figure 11 in Appendix G, suggesting that both sources tap into a shared internal component mediating the memorization–reasoning tradeoff, rather than dataset-specific artifacts.

Table 6 reports best-layer gains, macro-averaged across languages within each resource tier (to control for tier size) and then averaged over all five tasks. Full per-task Δ-accuracy grids, direction norms, and flip-rate diagnostics are provided in Appendix G.

6 Conclusion

We introduce **MIDI**, a multilingual idiom understanding benchmark covering 18 languages and dialects across high-, medium-, and low-resource tiers, with idioms presented in sentence and dialogue contexts and, where possible, in both figurative and literal forms. Zero-shot evaluations reveal that even state-of-the-art models struggle with idiom comprehension, with performance dropping sharply in low-resource languages and literal interpretations proving harder than figurative ones. Conversational context improves accuracy, particularly in low-resource settings, but does not fully close these gaps, highlighting persistent limitations in multilingual idiom understanding.

Our controlled probes show that reasoning with explicit definitions improves accuracy but does not fully close the gap. Memorization strongly correlates with successful figurative interpretation. Together, these findings suggest a hybrid view: memorization supplies candidate senses and priors, while reasoning leverages context to select the correct meaning. This explains why memorization aids reasoning and why these capacities are intertwined. Consistent with this entanglement, models exhibit a strong figurative bias, where priors boosting figurative accuracy can hinder literal comprehension when compositional interpretation is needed.

We further show that activation steering along memorization and reasoning dimensions yields modest yet consistent gains, especially in low-resource languages. Directions from MMLU-Pro transfer effectively to MIDI, and optimal layer distributions overlap, indicating a shared mechanism behind the memorization–reasoning tradeoff rather than dataset-specific artifacts.

Limitations

Scope of memorization vs. reasoning analysis. Our controlled experiments, which aim to disentangle memorization from contextual reasoning, focus only on *figurative* interpretations. The reasoning task asks, "What is the figurative meaning of [IDIOM]?", while the memorization task directly provides the figurative definition. Thus, our framework does not explicitly assess how memorization and reasoning interact for *literal* readings, where the model must suppress or override the usually preferred idiomatic meaning.

Uneven Coverage and Comparability Across Languages. Languages differ in whether idioms admit plausible literal counterparts (e.g., Vietnamese idioms in our dataset are purely figurative). Moreover, **MIDI** is imbalanced across resource tiers (12 low-resource vs. 3 medium and 3 high) and in the number of idioms/contexts per language. Thus, despite its size and diversity, **MIDI** is better suited for identifying broad tier-level and literal–figurative patterns than for tightly controlled, language-matched comparisons.

Agreement for Psycholinguistic Ratings. Although each instance was reviewed by at least two native speakers to verify linguistic quality and answer correctness, but we do not yet report inter-annotator agreement for familiarity, decomposability, or literal plausibility. These additional ratings should therefore be treated as informative but potentially noisy, rather than as definitive gold-standard measures.

Dialogue naturalness. Dialogue contexts are first generated by an LLM and then manually revised by native speakers to ensure semantic and linguistic accuracy. This yields coherent, well-controlled conversations but may underrepresent the stylistic and pragmatic variability found in naturally occurring interactions, slightly limiting naturalistic coverage.

MCQ Evaluation Constraints. We evaluate idiom comprehension using multiple-choice questions with manually curated distractors, allowing controlled comparisons across languages and settings. However, MCQs cannot fully capture open-ended interpretation or generation, so extending **MIDI** to free-form explanations and downstream tasks is left for future work.

Ethical Considerations

All human-authored data were manually reviewed to ensure the absence of harmful, offensive, or inappropriate content. Annotators provided informed consent for their contributions to be used and distributed for research purposes and were compensated through co-authorship. No sensitive or personally identifiable information was collected or disclosed, the dataset does not involve vulnerable populations, and the study poses minimal risk to participants.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of memory and language*, 27(6):668–683.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. 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.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Yihuai Hong, Dian Zhou, Meng Cao, Lei Yu, and Zhijing Jin. 2025. The reasoning-memorization interplay in language models is mediated by a single direction. *Preprint*, arXiv:2503.23084.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Gabriella Kovács. 2016. About the definition, classification, and translation strategies of idioms. *Acta Universitatis Sapientiae, Philologica*, 8(3):85–101.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

Maya R. Libben and Debra A. Titone. 2008. The multidetermined nature of idiom processing. *Memory & Cognition*, 36(6):1103–1121.

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.

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).

OpenAI. 2025. Gpt-5 system card.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jianing Zhou and Suma Bhat. 2024. Non-compositional expression generation and its continual learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2828–2839, Bangkok, Thailand. Association for Computational Linguistics.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2024. Enhancing language models with idiomatic reasoning. In *First Conference on Language Modeling*.

A Data Statements for MIDI

A.1 General Information

Dataset title: MIDI
Dataset curators: Not listed due to double-blind review.
Dataset version: 1.0 (November 2025)
Data statement authors: Not listed due to double-blind review.
Data statement version: 1.0 (December 2025)

A.2 Executive Summary

MIDI is a multilingual dataset for evaluating idiom understanding across languages with different resource levels. It covers 18 languages and dialects across high-, mid-, and low-resource tiers and contains 2,278 idioms instantiated in 8,460 usage contexts. Each idiom appears in both sentence-level and dialogue-level contexts and is paired with a multiple-choice question designed to test idiom comprehension under realistic discourse conditions. When a literal counterpart exists, both figurative and literal usages are included, with an average literal coverage of 82% across languages. The dataset supports controlled evaluation of idiom comprehension as well as diagnostic analyses that isolate memorization and reasoning using consistent prompt templates (Appendix C).

A.3 Curation Rationale

MIDI was curated to enable systematic analysis of idiom understanding in multilingual settings, particularly for languages that are underrepresented in existing benchmarks. Prior datasets often focus on a small number of high-resource languages or evaluate idioms in isolation, which limits their ability to capture how idioms are interpreted in context. MIDI addresses this gap by including a broad set of languages and by embedding idioms in both sentences and short dialogues. The inclusion of both figurative and literal usages further allows evaluation of how models distinguish idiomatic meaning from compositional interpretations.

A.4 Documentation for Source

Idioms were collected by native-speaker annotators for each language based on their linguistic knowledge and familiarity with common usage. The dataset does not rely on a single predefined lexicon or corpus. Example sentences were manually authored or curated to illustrate natural idiom usage. Dialogue contexts were initially generated using an LLM and then carefully reviewed and edited by native speakers to ensure linguistic quality and semantic correctness. All content was created specifically for this dataset to reduce the risk of data leakage and to maintain consistency across languages.

A.5 Language Variety

The dataset includes 18 languages and dialects spanning multiple language families, scripts, and typological properties. Languages are grouped into high-resource (Chinese, Japanese, Russian), mid-resource (Indonesian, Vietnamese, and the UAE dialect of Arabic), and low-resource categories (Javanese, Persian, Kannada, Telugu, Tamil, Minangkabau, Sundanese, Kazakh, Yoruba, and Arabic dialects from Syria, Egypt, and Morocco). Both standard and widely used regional varieties are included where appropriate, allowing analysis across different levels of linguistic resource availability.

B Dialogue Generation Prompts

GPT-4.5 prompt for figurative dialogue generation

Idiom: [IDIOM]
Idiom meaning: [IDIOM MEANING IN ENGLISH]
Generate a short 3-turn dialogue in which the final utterance includes the idiom "[IDIOM]" in [LANGUAGE]. After the dialogue, include the question: "Does [IDIOM] imply a figurative or literal meaning? What does it mean?"
Provide a multiple-choice question with three options. The correct answer is: [IDIOM MEANING IN NON-ENGLISH]. Create two incorrect options.
Output format (JSON):
{
"conversation": "The full 3-turn dialogue",
"question": "Does [IDIOM] imply a figurative or literal meaning? What does it mean?",
"correct option": "[IDIOM MEANING IN NON-ENGLISH]",
"incorrect option 1": "The idiom literal meaning in non English",
"incorrect option 2": "The other incorrect meaning with regard to its figurative meaning",
"incorrect option 3": "The other incorrect meaning with regard to its literal meaning"
}

Figure 4: Prompt used to generate dialogues where the idiom is intended to be interpreted figuratively.

GPT-4.5 prompt for literal dialogue generation

Phrase: [IDIOM]

Generate a short 3-turn dialogue in which the final utterance includes the phrase "[IDIOM]" in [LANGUAGE], used with its literal meaning. After the dialogue, include the question: "Does [IDIOM] imply a figurative or literal meaning? What does it mean?"

Provide a multiple-choice question with three options. The correct answer is: [IDIOM LITERAL MEANING IN NON-ENGLISH]. Create two incorrect options.

Output format (JSON):
{

"conversation": "The full 3-turn dialogue",
"question": "Does [IDIOM] imply a figurative or literal meaning? What does it mean?",
"correct option": "[IDIOM LITERAL MEANING IN NON-ENGLISH]",
"incorrect option 1": "[IDIOM FIGURATIVE MEANING IN NON-ENGLISH]",
"incorrect option 2": "The other incorrect meaning with regard to its figurative meaning",
"incorrect option 3": "The other incorrect meaning with regard to its literal meaning"
}

Figure 5: Prompt used to generate dialogues where the idiom is intended to be interpreted literally.

C Prompt Templates

This appendix lists the exact prompt templates used across all idiom comprehension settings. All prompts instantiate the placeholders {context}, {idiom}, {options}, and (when applicable) {idiom_meaning}.

C.1 Default Idiom Comprehension task (Reasoning + Memorization)

This is our main evaluation setting used in Section 4.2. The model receives the usage context (sentence or dialogue) and selects one out of four options: the figurative meaning, the literal meaning, and two distractors (one figurative, one literal). The correct choice depends on how the idiom usage type: if it is used literally, the answer reflects its direct, dictionary meaning; if it is used figuratively, the answer reflects its non-literal, symbolic meaning.

Reasoning + Memorization Prompt

You are tasked with selecting the most appropriate option based on the context provided below.
Context: {context}
What does the phrase "{idiom}" mean?
Options: options

C.2 Reasoning Isolation (Context + Meaning Hint)

To reduce reliance on definitional recall, we provide the idiom's meaning as a hint and evaluate whether models still select the correct answer conditioned on discourse context. This is used in reasoning task in Section 5.1.

Reasoning Isolation Prompt

You are tasked with selecting the most appropriate option based on the context provided below.
Context: {context}
"{idiom}" means {idiom_meaning}
What does the phrase "{idiom}" mean?
Options: {options}

C.3 Memorization Isolation (No Context)

This setting removes all usage context, probing definitional knowledge of idioms. The model must select the figurative meaning from the same four options. This is used in memorization task in Section 5.1.

Memorization Isolation Prompt

You are tasked with selecting the most appropriate option based on the context provided below.
What is the figurative meaning of "{idiom}"?
Options: {options}

C.4 Idiom Interpretation Bias

We measure interpretation bias using a binary choice format (AB). In this setup, {options} contains exactly two correct answers (e.g., figurative vs. literal), formatted as choices A and B. This is used in bias analysis in Section 5.2.

Model	Sentence			Conversation			Overall
	High	Mid	Low	High	Mid	Low	
Random	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Proprietary							
GPT-5.2	65.1	**90.0**	68.5	67.1	**91.1**	77.1	74.4
Gemini 2.5 Pro	**67.5**	89.5	**69.5**	71.2	90.9	**78.2**	**75.5**
Open-Source							
DeepSeek-R1 (70B)	48.8	53.6	34.6	54.2	57.5	40.5	44.0
Gemma-3 (27B)	**67.5**	86.9	58.8	**73.2**	90.0	70.9	71.0
Llama-3.1 (8B)	62.4	27.6	39.6	68.6	31.7	45.0	49.5
Llama-3.3 (70B)	64.6	65.5	53.5	65.2	75.2	62.4	63.6
Mixtral (8×7B)	69.4	59.6	42.6	76.9	75.6	53.2	57.5
Qwen-3 (4B)	30.6	31.0	27.5	38.3	35.1	28.5	30.8
Qwen-3 (30B-MoE)	40.8	24.1	25.6	37.4	23.5	25.8	31.2
Average	*57.4*	*58.6*	*46.7*	*61.4*	*63.4*	*53.5*	*55.3*

High = *High Resource* *Mid* = *Mid Resource* *Low* = *Low Resource*

Table 7: Model performance (accuracy) on idiom comprehension by usage context (Sentence vs. Conversation), stratified by language resource availability. **Bold** indicates best; underline indicates second-best.

Model	Figurative			Literal			Overall
	High	Mid	Low	High	Mid	Low	
Random	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Proprietary							
GPT-5.2	97.0	97.2	86.7	35.2	76.8	58.9	74.4
Gemini 2.5 Pro	**98.6**	**99.2**	**94.8**	40.2	72.1	52.9	**75.5**
Open-Source							
DeepSeek-R1 (70B)	55.1	57.5	42.6	47.9	46.4	32.6	44.0
Gemma-3 (27B)	96.9	96.2	76.4	43.8	71.7	53.3	71.0
Llama-3.1 (8B)	82.0	29.3	43.1	49.0	33.5	41.6	49.5
Llama-3.3 (70B)	87.5	77.9	70.7	42.3	59.4	45.2	63.6
Mixtral (8×7B)	79.4	57.9	35.7	**67.0**	**80.7**	**60.1**	57.5
Qwen-3 (4B)	43.3	36.3	27.1	25.7	32.8	28.9	30.8
Qwen-3 (30B-MoE)	43.9	24.4	24.1	34.3	23.3	27.2	31.2
Average	*76.0*	*64.0*	*55.7*	*42.8*	*55.2*	*44.5*	*55.3*

High = *High Resource* *Mid* = *Mid Resource* *Low* = *Low Resource*

Table 8: Model performance (accuracy) on idiom comprehension by usage type (Figurative vs. Literal), stratified by language resource availability. **Bold** indicates best; underline indicates second-best.

Bias Prompt (Binary A/B)

You are tasked with selecting the most appropriate option based on the context provided below. What does the phrase "{idiom}" mean? Options: {options}

Option formatting. For the 4-way MCQ settings, {options} is formatted as four labeled choices (A–D). For the bias setting, {options} is formatted as two labeled choices (A–B). We keep the question surface form fixed across settings whenever possible to avoid confounding effects from prompt formulation.

D Detailed Main Results

In addition to the main results presented in the paper, Tables 7 and 8 provide detailed breakdowns of model performance.

E Reasoning Evaluation Details

Sentence vs. conversation in reasoning-only. Table 9 breaks down the reasoning-only setting (Appendix C.2) by context type. Differences between sentence and dialogue contexts are small across models and resource tiers (typically within 1–2 points on average), with proprietary models remaining near ceiling in both settings. This suggests that once the idiom's English meaning is provided, performance is largely insensitive to whether the usage context is a single sentence or a short dialogue.

Model	Sentence			Conversation			Overall
	High	Mid	Low	High	Mid	Low	
Random	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*	*25.0*
Proprietary							
GPT-5.2	98.4	98.7	94.7	98.0	98.3	96.4	96.4
Gemini 2.5 Pro	**99.1**	**99.7**	**96.5**	**98.8**	**99.7**	**98.5**	**98.0**
Open-Source							
DeepSeek-R1 (70B)	53.0	59.6	52.0	55.6	63.0	55.8	54.7
Gemma-3 (27B)	97.8	98.7	91.3	97.1	99.0	91.1	93.9
Llama-3.1 (8B)	97.4	55.4	72.8	96.2	53.2	74.5	78.1
Llama-3.3 (70B)	96.2	88.2	76.4	95.6	93.8	79.3	85.2
Mixtral (8×7B)	94.0	89.9	72.6	95.1	92.2	71.8	80.8
Qwen-3 (4B)	44.7	38.5	30.7	44.0	35.0	32.8	36.9
Qwen-3 (30B-MoE)	66.9	35.6	30.5	64.1	36.1	32.1	43.0
Average	*83.1*	*73.8*	*68.6*	*82.7*	*74.5*	*70.3*	*74.1*

High = *High Resource* *Mid* = *Mid Resource* *Low* = *Low Resource*

Table 9: Model performance (accuracy) on idiom reasoning by usage context (Sentence vs. Conversation), evaluating figurative meaning identification when provided with usage context and English meaning. Stratified by language resource availability. **Bold** indicates best; underline indicates second-best.

F Interpretation Bias Details

When presented with ambiguous idiom prompts lacking contextual cues ("What does the phrase *{idiom}* mean?"), models exhibit systematic interpretation biases. Table 10 reports interpretation tendencies alongside weighted overall accuracy on figurative (n=2,278) and literal (n=1,952) meaning identification tasks, i.e Section 4.2.

Figures 6 and 7 visualize interpretation bias sorted by task performance to examine potential correlations. Models with stronger figurative bias generally achieve higher figurative accuracy (Figure 6), suggesting that default figurative interpretation aligns with task demands. Notably, Mixtral achieves the highest literal accuracy despite near-

Model	Interpretation (%) Fig.	Interpretation (%) Lit.	Interpretation (%) Δ	Overall Accuracy Fig.	Overall Accuracy Lit.
Random	*50.0*	*50.0*	*0.0*	*25.0*	*25.0*
Proprietary					
Gemini 2.5 Pro	**89.7**	10.3	**79.4**	**96.4**	50.8
GPT-5.2	77.4	22.6	54.9	90.9	53.2
Open-Source					
DeepSeek-R1 (70B)	47.5	52.5	-5.0	48.0	38.8
Gemma-3 (27B)	78.7	21.3	57.4	84.7	52.1
Llama-3.1 (8B)	53.9	46.1	7.8	52.0	43.1
Llama-3.3 (70B)	66.1	33.9	32.3	76.3	45.7
Mixtral (8×7B)	48.9	**51.1**	-2.1	50.8	**64.4**
Qwen-3 (4B)	66.7	33.3	33.4	32.8	28.3
Qwen-3 (30B-MoE)	57.4	42.6	14.7	29.6	29.1
Average	*65.2*	*34.8*	*30.3*	*62.4*	*45.1*

Δ = Figurative – Literal bias; Fig. = Figurative, Lit. = Literal

Table 10: Model interpretation bias and reasoning accuracy on ambiguous idiom prompts. *Left*: Interpretation tendency when no context is provided (positive Δ = figurative preference). *Right*: Weighted overall accuracy on figurative (n=2,278) and literal (n=1,952) meaning identification. **Bold** = highest; underline = second-highest.

neutral bias, indicating that balanced interpretation may benefit literal comprehension (Figure 7).

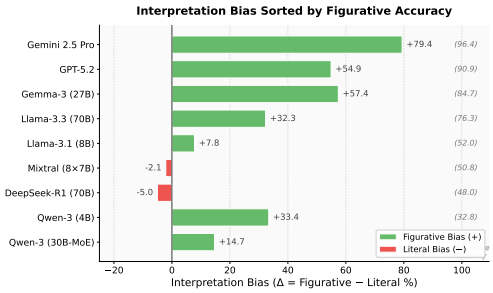

Figure 6: Interpretation bias (Δ) sorted by figurative accuracy. Positive Δ indicates figurative preference. Scores in parentheses denote weighted accuracy.

G Activation Steering Details

This appendix provides the full steering diagnostics referenced in Section 5.3: per-task Δ-accuracy grids, layer-sweep statistics, direction magnitudes, and flip-rate analysis.

G.1 Additional Steering Definitions

For auxiliary analyses we quantify how well an input aligns with the memorization→reasoning direction using a normalized projection score (Hong et al., 2025):

$$\hat{r}^{(\ell)} = \frac{r^{(\ell)}}{\|r^{(\ell)}\|_2}, \quad s^{(\ell)}(x) = \hat{r}^{(\ell)\top} h^{(\ell)}(x). \quad (3)$$

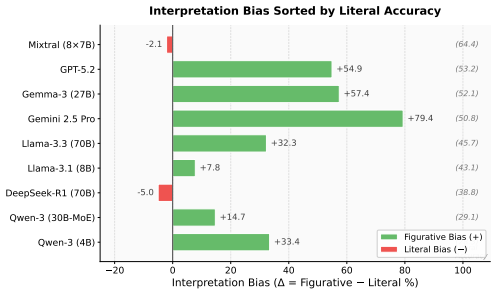

Figure 7: Interpretation bias (Δ) sorted by literal accuracy. Models with lower figurative bias (e.g., Mixtral) tend to perform better on literal interpretation.

We also define an *ablation* operator that removes activation components along the reasoning direction (not used in our primary experiments):

$$h^{(\ell)}(x) \leftarrow h^{(\ell)}(x) - \hat{r}^{(\ell)} \left(\hat{r}^{(\ell)\top} h^{(\ell)}(x) \right). \quad (4)$$

G.2 Where Steering Helps

Figures 8 and 9 report steering-induced accuracy changes (in percentage points), stratified by resource tier. A consistent pattern emerges across tasks: **low-resource languages exhibit the largest average gains**, while high-resource languages show more modest changes. This aligns with a natural *headroom* effect, as high-resource settings often approach ceiling performance, leaving limited room for improvement.

G.3 Layer Sensitivity and Direction Magnitude

Steering effectiveness is **highly layer-dependent**: optimal layers vary across languages and tasks. Figure 11 shows the distribution of best-performing layers from our sweep. Within each model, the best-layer distributions for **MMLU-Pro** and **MIDI-derived** vectors are strikingly similar, supporting the hypothesis that both probe a shared internal component related to the memorization–reasoning tradeoff (rather than dataset-specific artifacts).

For both models, layer 3 emerges as the most frequently optimal choice. Median best layers cluster at 3–4 for Llama-3.1 8B (median of 4 with MMLU-Pro vectors; median of 3 with MIDI-derived vectors), whereas Qwen-3 4B exhibits a median of 7 for both vector sources, reflecting a broader distribution toward mid-depth layers.

Notably, suboptimal layer choices can be actively harmful. Across all language and task configurations, the *worst* layer in each sweep reduces accuracy by −0.88 points on average, with the most

Model / Setting	Sentence (Accuracy %)			Conversation (Accuracy %)		
	High	Mid	Low	High	Mid	Low
Reasoning + Memorization						
Qwen-3 (4B)	61.86	71.82	44.90	67.84	59.66	38.31
+ memorization vect	62.89 (+1.03)	72.49 (+0.67)	46.24 (+1.34)	**68.63** (+0.79)	59.83 (+0.17)	40.04 (+1.73)
+ reasoning vect	62.82 (+0.96)	72.49 (+0.67)	46.44 (+1.54)	68.55 (+0.71)	60.33 (+0.67)	39.96 (+1.65)
Llama-3.1 (8B)	63.48	78.63	52.84	62.17	77.59	51.77
+ memorization vect	**63.97** (+0.49)	**79.50** (+0.87)	53.30 (+0.46)	62.61 (+0.44)	77.59 (+0.00)	52.22 (+0.45)
+ reasoning vect	63.92 (+0.44)	**79.50** (+0.87)	**53.36** (+0.52)	62.54 (+0.37)	**77.75** (+0.16)	**52.37** (+0.60)
Reasoning						
Qwen-3 (4B)	97.93	**98.33**	83.32	**98.11**	98.00	83.73
+ memorization vect	97.93 (+0.00)	**98.33** (+0.00)	83.90 (+0.58)	**98.11** (+0.00)	**98.67** (+0.67)	84.06 (+0.33)
+ reasoning vect	**98.04** (+0.11)	**98.33** (+0.00)	83.55 (+0.23)	**98.11** (+0.00)	98.33 (+0.33)	84.25 (+0.51)
Llama-3.1 (8B)	97.62	97.69	87.29	96.00	97.36	86.05
+ memorization vect	97.62 (+0.00)	98.02 (+0.33)	87.41 (+0.12)	96.00 (+0.00)	97.36 (+0.00)	**86.42** (+0.38)
+ reasoning vect	97.62 (+0.00)	97.69 (+0.00)	**87.51** (+0.22)	96.11 (+0.11)	97.36 (+0.00)	86.17 (+0.13)
Memorization						
Qwen-3 (4B)		—		86.62	66.22	43.37
+ memorization vect		—		**88.03** (+1.41)	66.53 (+0.31)	45.00 (+1.64)
+ reasoning vect		—		87.58 (+0.74)	**66.86** (+0.64)	44.80 (+1.35)
Llama-3.1 (8B)		—		85.34	62.25	44.14
+ memorization vect		—		85.45 (+0.11)	62.56 (+0.31)	**45.62** (+1.48)
+ reasoning vect		—		85.34 (+0.00)	63.20 (+0.95)	45.37 (+1.23)

Table 11: **MMLU-Pro steering vectors** applied to to both conversation and sentence tasks. Values are language accuracies (%) aggregated over high-/mid-/low-resource languages under sentence and conversation evaluation. Parentheses denote absolute change vs. the unsteered baseline for the same model and split. **Bold** marks the best score across both models (within each evaluation split and column).

severe case yielding a -6.67 point drop. This underscores that steering should be approached as a calibrated intervention rather than assumed to yield unconditional improvements.

Figure 10 plots the ℓ_2 norms of direction vectors across layers. When averaged across layers (excluding layer 0, where some directions are exactly zero), Qwen's direction vectors exhibit substantially larger norms than Llama's: approximately $5.0\times$ larger for MIDI-derived vectors and $7.4\times$ larger for MMLU-Pro vectors. This disparity is consistent with Qwen's heightened behavioral sensitivity to steering at equivalent scales.

G.4 Flip-Rate Analysis

To quantify how frequently steering alters model predictions, we compute **flip rates** on the complete evaluation set: the proportion of examples where the steered prediction differs from the baseline. We disaggregate these into *improvement flips* (baseline incorrect $\rightarrow$ steered correct) and *regression flips* (baseline correct $\rightarrow$ steered incorrect).

Figure 12 presents flip rates by resource tier, averaged across steering types. Flip rates remain modest in absolute terms (typically below a few percent), but are systematically elevated in low-resource settings, paralleling the larger accuracy gains in Figures 8 and 9.

H Dataset Examples

Table 16 presents example instances from the **MIDI** dataset. In addition, Table 17 reports descriptive statistics for both idiomatic and literal expressions in sentence- and dialogue-level contexts.

Model / Setting	Sentence (Accuracy %)			Conversation (Accuracy %)		
	High	Mid	Low	High	Mid	Low
Reasoning + Memorization						
Qwen-3 (4B)	61.86	71.82	44.90	67.84	59.66	38.31
+ memorization vect	62.56 (+0.70)	72.49 (+0.67)	46.32 (+1.42)	**68.67** (+0.83)	60.00 (+0.34)	39.51 (+1.19)
+ reasoning vect	62.83 (+0.97)	72.20 (+0.38)	46.25 (+1.35)	68.43 (+0.59)	60.00 (+0.34)	39.26 (+0.95)
Llama-3.1 (8B)	63.48	78.63	52.84	62.17	77.59	51.77
+ memorization vect	**63.97** (+0.49)	**79.50** (+0.87)	**53.36** (+0.52)	62.83 (+0.66)	**77.75** (+0.16)	**52.38** (+0.61)
+ reasoning vect	63.82 (+0.34)	79.33 (+0.70)	53.28 (+0.44)	62.64 (+0.47)	**77.75** (+0.16)	52.33 (+0.56)
Reasoning						
Qwen-3 (4B)	97.93	**98.33**	83.32	**98.11**	98.00	83.73
+ memorization vect	97.93 (+0.00)	**98.33** (+0.00)	83.90 (+0.58)	98.00 (-0.11)	**98.33** (+0.33)	84.12 (+0.39)
+ reasoning vect	**98.04** (+0.11)	**98.33** (+0.00)	83.55 (+0.23)	**98.11** (+0.00)	**98.33** (+0.33)	84.12 (+0.39)
Llama-3.1 (8B)	97.62	97.69	87.29	96.00	97.36	86.05
+ memorization vect	**97.62** (+0.00)	98.02 (+0.33)	87.41 (+0.12)	96.11 (+0.11)	97.36 (+0.00)	86.07 (+0.03)
+ reasoning vect	**97.62** (+0.00)	97.69 (+0.00)	**87.51** (+0.22)	96.11 (+0.11)	97.36 (+0.00)	**86.17** (+0.13)
Memorization						
Qwen-3 (4B)		—		86.62	66.22	43.37
+ memorization vect		—		**87.58** (+0.96)	**66.53** (+0.31)	44.53 (+1.16)
+ reasoning vect		—		87.32 (+0.69)	**66.53** (+0.31)	44.78 (+1.42)
Llama-3.1 (8B)		—		85.34	62.25	44.14
+ memorization vect		—		85.34 (+0.00)	63.22 (+0.98)	45.51 (+1.37)
+ reasoning vect		—		85.45 (+0.11)	62.86 (+0.62)	**45.58** (+1.44)

Table 12: **MIDI steering vectors** for separating memorization and reasoning. Values are language accuracies (%) aggregated over high-/mid-/low-resource languages under sentence and conversation evaluation. Parentheses denote absolute change vs. the unsteered baseline for the same model and split. **Bold** marks the best score across both models (within each evaluation split and column).

Qwen-3 (4B)					Llama-3.1 (8B)				
Language	Resource class.	Base	+Mem	+Reas	Language	Resource class.	Base	+Mem	+Reas
Chinese	High	68.00	68.50 (+0.50)	68.50 (+0.50)	Chinese	High	59.17	59.33 (+0.16)	59.83 (+0.66)
Japanese	High	71.88	72.32 (+0.44)	72.77 (+0.89)	Japanese	High	66.07	66.96 (+0.89)	66.52 (+0.45)
Russian	High	63.66	65.08 (+1.42)	64.37 (+0.71)	Russian	High	61.28	61.52 (+0.24)	61.28 (+0.00)
Arabic_UAE	Mid	23.23	23.74 (+0.51)	23.23 (+0.00)	Arabic_UAE	Mid	60.61	60.61 (+0.00)	61.11 (+0.50)
Indonesia	Mid	75.76	75.76 (+0.00)	75.76 (+0.00)	Indonesia	Mid	75.15	75.15 (+0.00)	75.15 (+0.00)
Vietnam	Mid	80.00	80.00 (+0.00)	82.00 (+2.00)	Vietnam	Mid	97.00	97.00 (+0.00)	97.00 (+0.00)
Arabic_Egypt	Low	29.35	31.52 (+2.17)	31.52 (+2.17)	Arabic_Egypt	Low	59.24	59.24 (+0.00)	59.24 (+0.00)
Arabic_Morocco	Low	25.39	26.42 (+1.03)	26.42 (+1.03)	Arabic_Morocco	Low	57.51	57.51 (+0.00)	57.51 (+0.00)
Arabic_Syrian	Low	25.00	27.00 (+2.00)	26.50 (+1.50)	Arabic_Syrian	Low	55.50	56.00 (+0.50)	56.50 (+1.00)
Persian	Low	51.09	51.63 (+0.54)	52.72 (+1.63)	Persian	Low	65.22	65.22 (+0.00)	65.76 (+0.54)
Javanese	Low	43.00	45.50 (+2.50)	45.50 (+2.50)	Javanese	Low	54.00	54.50 (+0.50)	54.50 (+0.50)
Kannada	Low	61.56	63.32 (+1.76)	63.82 (+2.26)	Kannada	Low	67.34	67.84 (+0.50)	68.34 (+1.00)
Kazakh	Low	46.50	47.50 (+1.00)	47.50 (+1.00)	Kazakh	Low	45.00	46.50 (+1.50)	46.50 (+1.50)
Minangkabau	Low	27.52	30.87 (+3.35)	30.20 (+2.68)	Minangkabau	Low	33.56	34.23 (+0.67)	34.23 (+0.67)
Sundanese	Low	27.91	30.23 (+2.32)	30.23 (+2.32)	Sundanese	Low	45.93	45.93 (+0.00)	46.51 (+0.58)
Tamil	Low	58.67	60.20 (+1.53)	59.18 (+0.51)	Tamil	Low	60.71	61.73 (+1.02)	61.73 (+1.02)
Telugu	Low	39.13	40.94 (+1.81)	40.58 (+1.45)	Telugu	Low	48.91	49.64 (+0.73)	49.28 (+0.37)
Yoruba	Low	24.63	25.37 (+0.74)	25.37 (+0.74)	Yoruba	Low	28.36	28.36 (+0.00)	28.36 (+0.00)

Table 13: **Conversation Vanilla Task.** Per-language (country) results for **MMLU-Pro vector steering**. We report accuracy (%) for the baseline model (**Base**) and after applying memorization (**+Mem**) or reasoning (**+Reas**) steering; parentheses denote the absolute change vs. baseline. For each model and steering type, the intervention is applied at the **best-performing layer** (selected on a validation set). **Resource classification** indicates language resource level (**High** , **Mid** , **Low**).

Qwen-3 (4B)					Llama-3.1 (8B)				
Language	**Resource class.**	**Base**	**+Mem**	**+Reas**	**Language**	**Resource class.**	**Base**	**+Mem**	**+Reas**
Chinese	**High**	95.67	95.67 (+0.00)	95.67 (+0.00)	Chinese	**High**	90.67	90.67 (+0.00)	91.00 (+0.33)
Japanese	**High**	99.13	99.13 (+0.00)	99.13 (+0.00)	Japanese	**High**	98.26	98.26 (+0.00)	98.26 (+0.00)
Russian	**High**	99.53	99.53 (+0.00)	99.53 (+0.00)	Russian	**High**	99.06	99.06 (+0.00)	99.06 (+0.00)
Arabic_UAE	**Mid**	96.00	97.00 (+1.00)	96.00 (+0.00)	Arabic_UAE	**Mid**	97.00	97.00 (+0.00)	97.00 (+0.00)
Indonesia	**Mid**	100.00	100.00 (+0.00)	100.00 (+0.00)	Indonesia	**Mid**	99.07	99.07 (+0.00)	99.07 (+0.00)
Vietnam	**Mid**	98.00	99.00 (+1.00)	99.00 (+1.00)	Vietnam	**Mid**	96.00	96.00 (+0.00)	96.00 (+0.00)
Arabic_Egypt	**Low**	98.00	99.00 (+1.00)	99.00 (+1.00)	Arabic_Egypt	**Low**	98.00	98.00 (+0.00)	98.00 (+0.00)
Arabic_Morocco	**Low**	95.92	95.92 (+0.00)	95.92 (+0.00)	Arabic_Morocco	**Low**	94.90	95.92 (+1.02)	95.92 (+1.02)
Arabic_Syrian	**Low**	99.00	99.00 (+0.00)	99.00 (+0.00)	Arabic_Syrian	**Low**	99.00	99.00 (+0.00)	99.00 (+0.00)
Persian	**Low**	94.12	96.08 (+1.96)	97.06 (+2.94)	Persian	**Low**	98.04	98.04 (+0.00)	98.04 (+0.00)
Javanese	**Low**	99.04	99.04 (+0.00)	99.04 (+0.00)	Javanese	**Low**	97.12	97.12 (+0.00)	97.12 (+0.00)
Kannada	**Low**	96.46	96.46 (+0.00)	95.96 (-0.50)	Kannada	**Low**	98.99	99.49 (+0.50)	99.49 (+0.50)
Kazakh	**Low**	85.00	85.00 (+0.00)	87.00 (+2.00)	Kazakh	**Low**	98.00	98.00 (+0.00)	98.00 (+0.00)
Minangkabau	**Low**	70.00	70.00 (+0.00)	69.00 (-1.00)	Minangkabau	**Low**	68.00	68.00 (+0.00)	68.00 (+0.00)
Sundanese	**Low**	74.00	74.00 (+0.00)	74.00 (+0.00)	Sundanese	**Low**	70.00	71.00 (+1.00)	70.00 (+0.00)
Tamil	**Low**	95.96	95.96 (+0.00)	95.96 (+0.00)	Tamil	**Low**	96.97	97.98 (+1.01)	96.97 (+0.00)
Telugu	**Low**	83.45	83.45 (+0.00)	84.17 (+0.72)	Telugu	**Low**	79.86	79.86 (+0.00)	79.86 (+0.00)
Yoruba	**Low**	13.86	14.85 (+0.99)	14.85 (+0.99)	Yoruba	**Low**	33.66	34.65 (+0.99)	33.66 (+0.00)

Table 14: **Reasoning task (Conversation).** Per-language (country) results with **MMLU-Pro vector steering** (same reporting as Table 13). Parentheses denote the absolute change vs. baseline; steering is applied at the **best-performing layer** selected on a validation set. **Resource classification** indicates language resource level (**High** , **Mid** , **Low**).

Qwen-3 (4B)					Llama-3.1 (8B)				
Language	**Resource class.**	**Base**	**+Mem**	**+Reas**	**Language**	**Resource class.**	**Base**	**+Mem**	**+Reas**
Chinese	**High**	99.33	100.00 (+0.67)	100.00 (+0.67)	Chinese	**High**	97.00	97.33 (+0.33)	97.00 (+0.00)
Japanese	**High**	86.96	89.57 (+2.61)	88.70 (+1.74)	Japanese	**High**	82.61	82.61 (+0.00)	82.61 (+0.00)
Russian	**High**	73.58	74.53 (+0.95)	74.06 (+0.48)	Russian	**High**	76.42	76.42 (+0.00)	76.42 (+0.00)
Arabic_UAE	**Mid**	53.00	53.00 (+0.00)	53.00 (+0.00)	Arabic_UAE	**Mid**	54.00	54.00 (+0.00)	55.00 (+1.00)
Indonesia	**Mid**	66.67	67.59 (+0.92)	67.59 (+0.92)	Indonesia	**Mid**	65.74	66.67 (+0.93)	67.59 (+1.85)
Vietnam	**Mid**	79.00	79.00 (+0.00)	80.00 (+1.00)	Vietnam	**Mid**	67.00	67.00 (+0.00)	67.00 (+0.00)
Arabic_Egypt	**Low**	57.00	59.00 (+2.00)	59.00 (+2.00)	Arabic_Egypt	**Low**	68.00	70.00 (+2.00)	70.00 (+2.00)
Arabic_Morocco	**Low**	59.79	60.82 (+1.03)	61.86 (+2.07)	Arabic_Morocco	**Low**	57.73	59.79 (+2.06)	59.79 (+2.06)
Arabic_Syrian	**Low**	54.00	53.00 (-1.00)	53.00 (-1.00)	Arabic_Syrian	**Low**	51.00	52.00 (+1.00)	52.00 (+1.00)
Persian	**Low**	44.12	46.08 (+1.96)	44.12 (+0.00)	Persian	**Low**	50.98	51.96 (+0.98)	51.96 (+0.98)
Javanese	**Low**	36.54	36.54 (+0.00)	36.54 (+0.00)	Javanese	**Low**	32.69	33.65 (+0.96)	32.69 (+0.00)
Kannada	**Low**	41.41	41.92 (+0.51)	42.42 (+1.01)	Kannada	**Low**	44.95	45.96 (+1.01)	44.95 (+0.00)
Kazakh	**Low**	46.00	50.00 (+4.00)	50.00 (+4.00)	Kazakh	**Low**	45.00	46.00 (+1.00)	45.00 (+0.00)
Minangkabau	**Low**	34.00	37.00 (+3.00)	37.00 (+3.00)	Minangkabau	**Low**	29.00	34.00 (+5.00)	34.00 (+5.00)
Sundanese	**Low**	26.00	26.00 (+0.00)	26.00 (+0.00)	Sundanese	**Low**	30.00	32.00 (+2.00)	31.00 (+1.00)
Tamil	**Low**	53.06	55.10 (+2.04)	53.06 (+0.00)	Tamil	**Low**	43.88	44.90 (+1.02)	44.90 (+1.02)
Telugu	**Low**	41.73	43.88 (+2.15)	43.88 (+2.15)	Telugu	**Low**	46.76	47.48 (+0.72)	47.48 (+0.72)
Yoruba	**Low**	26.73	30.69 (+3.96)	28.71 (+1.98)	Yoruba	**Low**	29.70	29.70 (+0.00)	30.69 (+0.99)

Table 15: **Memorization task.** Per-language (country) results with **MMLU-Pro vector steering** (same reporting as Table 13). Parentheses denote the absolute change vs. baseline; steering is applied at the **best-performing layer** selected on a validation set. **Resource classification** indicates language resource level (**High** , **Mid** , **Low**).

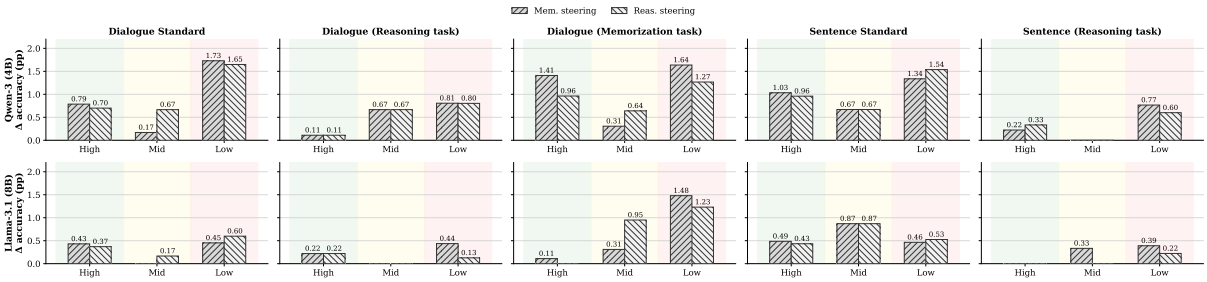

Figure 8: Steering-induced accuracy gains (Δ) using **MMLU-Pro vectors**, broken down by task and resource tier.

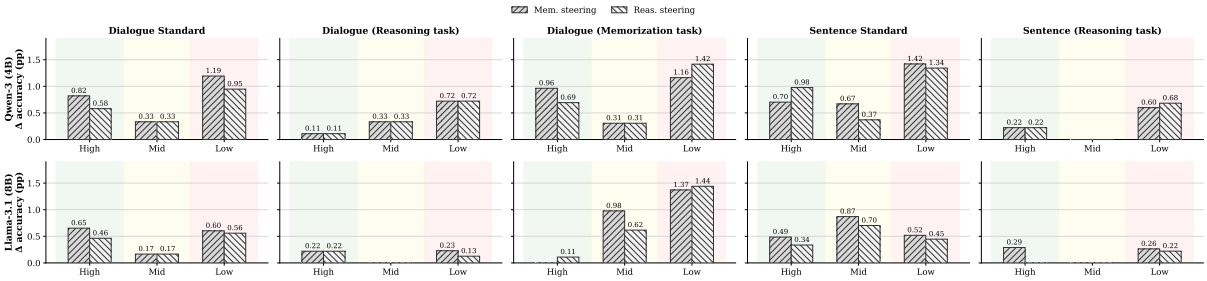

Figure 9: Steering-induced accuracy gains (Δ) using **MIDI-derived vectors**, broken down by task and resource tier.

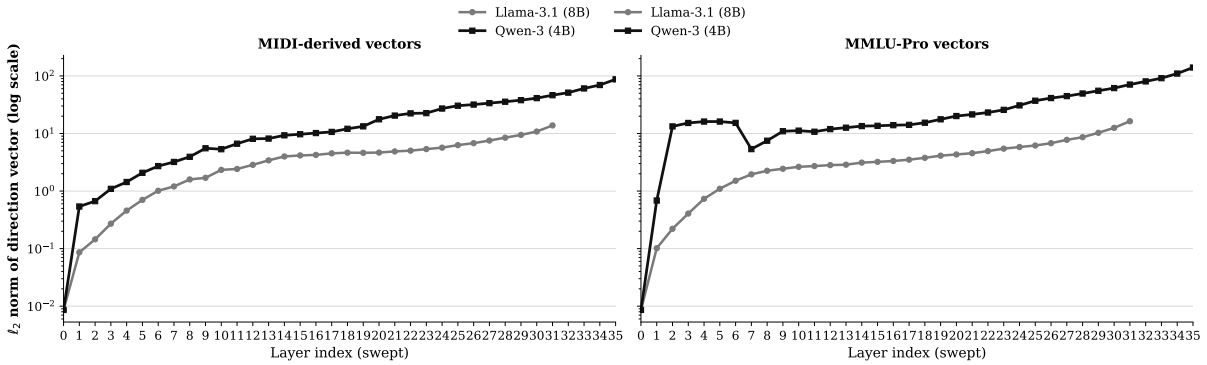

Figure 10: ℓ_2 norms of the memorization→reasoning direction vector across swept layers (log scale).

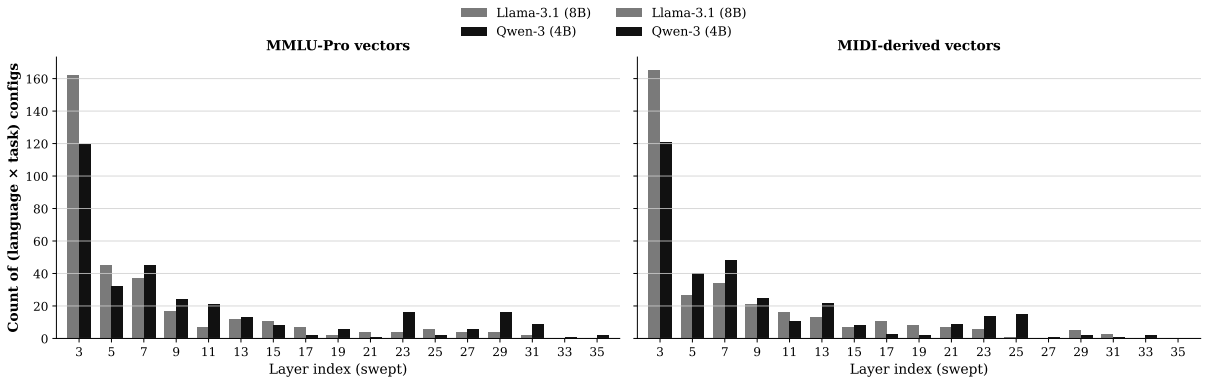

Figure 11: Distribution of best-performing layers (selected by highest development accuracy) from the layer sweep. We evaluate every other layer starting at ℓ=3 (Llama: through ℓ=31; Qwen: through ℓ=35).

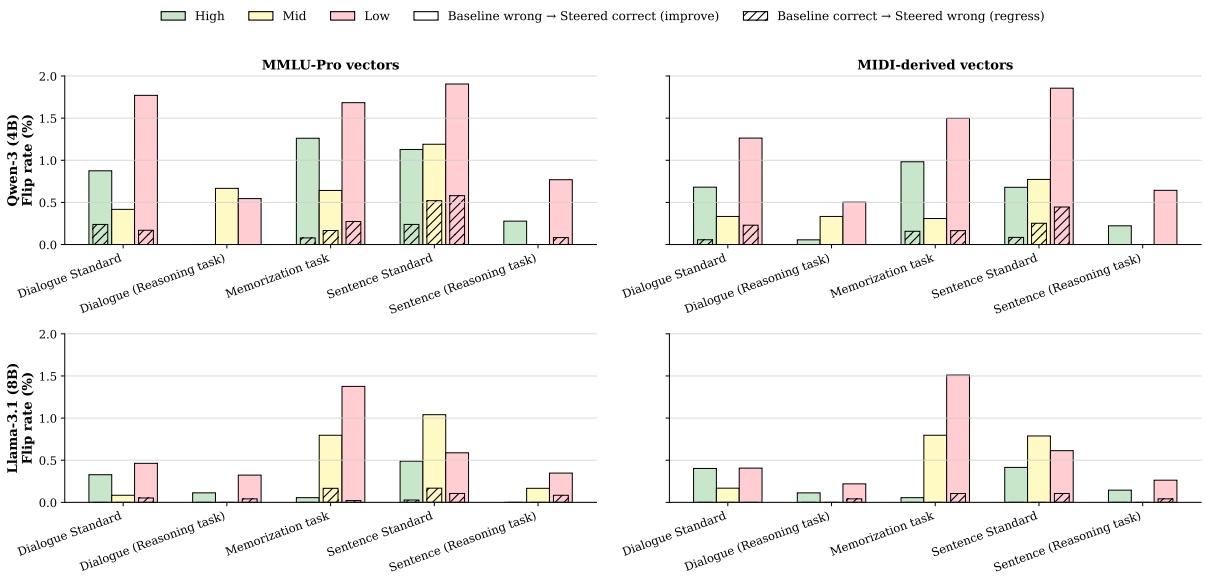

Figure 12: Flip rates on the full evaluation set, macro-averaged across languages within each resource tier and subsequently averaged over tasks. Improvement flips appear as solid bars; regression flips are indicated with hatched overlays.

Idiom	Definition	Definition Non-English	Type	Sentence	Dialogue	Questions	Choices	Answer Key
Bắt cá hai tay	Cunning, greedy action, want to have many things at one time	Hành động khôn lỏi, tham lam, muốn có nhiều thứ	figurative	Anh ấy bị phát hiện bắt cá hai tay nên cả hai cô gái đều chia tay anh ta.	A: Nghe nói Minh cùng lúc tán tỉnh hai người bạn cùng lớp? B: Đúng rồi, cậu ấy không quyết định nổi ai cả. A: Rõ là bắt cá hai tay.	What does the phrase Bắt cá hai tay mean?	A. Tài năng bắt được cá bằng cả hai tay B. Chỉ người chăm chỉ làm nhiều việc cùng lúc C. Bắt được hai con cá cùng một lúc D. Hành động khôn lỏi, muốn có nhiều thứ	D
ẹja n bakan?	is it positive or negative	so bọsi abi ko bọsi	literal	ṣe ẹja n bakan lo fe lori ounje re?	A : baba lagabja, booni elo ounje aleyi, ẹja abi akan? baba lagabja: ah eja ni o, ti mo ba ri ponmo na iyen na a lọ A: oda, olohun a ṣe iyanu	What does the phrase ẹja n bakan? mean?	A. se eja tabi akan ni won o fi jehun B. se baba lagbaja fẹran akan tabi ẹja? C. on bere pe se o lo eja tabi akan D. se oro ti won so bọsi tabi ko bọsi	A

Table 16: Example instances from the MIDI dataset.

Language	Figurative (Sentence)		Figurative (Dialogue)		Literal (Sentence)		Literal (Dialogue)	
	#Wrds	#Chrs	#Wrds	#Chrs	#Wrds	#Chrs	#Wrds	#Chrs
High-Resource Languages								
Chinese	17.12	25.67	40.83	61.19	32.99	47.54	49.07	72.62
Japanese	13.00	19.37	37.24	59.54	15.29	23.08	48.43	75.60
Russian	12.05	73.77	31.46	187.02	13.05	82.49	33.49	200.80
Mid-Resource Languages								
Arabic (UAE)	11.23	60.63	22.36	120.40	11.59	63.52	25.81	139.37
Indonesian	11.38	77.56	24.81	146.81	10.86	74.28	27.81	161.89
Vietnamese	19.72	85.86	36.63	159.38	11.00	58.00	7.00	41.00
Low-Resource Languages								
Arabic (Egypt)	8.78	43.15	22.35	117.36	7.44	37.07	25.88	137.64
Arabic (Morocco)	11.92	61.86	23.49	128.55	10.58	56.71	28.47	160.89
Arabic (Syria)	7.55	38.53	22.32	118.00	8.84	46.86	25.81	136.32
Persian	12.99	63.64	26.33	134.36	12.76	61.12	31.84	160.94
Javanese	13.45	85.06	28.22	162.50	10.80	67.71	29.20	168.42
Kannada	6.74	54.13	30.55	235.51	6.18	48.06	22.82	174.18
Kazakh	7.28	49.64	27.47	187.98	6.35	42.17	24.09	162.36
Minangkabau	8.42	50.23	30.19	159.67	9.47	57.24	29.61	161.41
Sundanese	13.53	79.28	25.48	150.36	13.12	77.53	28.11	163.88
Tamil	7.71	64.05	20.92	170.01	5.93	48.31	21.62	177.80
Telugu	7.16	53.66	15.37	93.73	6.15	46.58	19.68	113.92
Yoruba	12.50	52.35	26.77	115.20	11.20	57.87	26.61	115.24
Average	**11.25**	**57.69**	**27.38**	**139.31**	**11.31**	**55.34**	**28.08**	**140.24**

Table 17: Average length statistics (words and characters) for idiom usage across sentence and dialogue contexts, separated by figurative and literal interpretations and grouped by language resource level.