



From task performance to linguistic evidence: Non-literal language in the age of large language models

Derek F. Wong* · Fengying Ye

(University of Macau)

Wong, Derek F. and Fengying Ye. 2026. From task performance to linguistic evidence: Non-literal language in the age of large language models. *Linguistic Research* 43(Special Edition): 297-321. This paper examines non-literal language as a diagnostic site for rethinking linguistic evidence in the age of large language models (LLMs). As language models become increasingly capable of generating, comparing, and evaluating language at scale, their behavior can be analyzed through linguistically motivated criteria, rather than only through measures of task performance. We develop this argument through two analyses: metaphor understanding and cross-lingual idiom alignment. For metaphor, we investigate whether models can identify the attribute projected by a metaphorical expression, rather than being guided primarily by salient words or disrupted by changes in syntactic form. For idioms, we ask whether models can identify meaning-equivalent idioms across languages, rather than relying on literal translation, word overlap, or contextually plausible but non-equivalent expressions. Taken together, these analyses show that model behavior should be understood not merely as task performance, but as evidence of how fluent output relates to linguistically defined meaning. This perspective shifts attention from whether models produce correct answers to what their behavior reveals about meaning interpretation. We argue that these patterns point to a productive interaction between linguistics and AI: linguistic analysis provides the categories needed to define what counts as meaning, understanding, and evidence, while large-scale model behavior provides a new empirical setting in which these categories can be examined. (University of Macau)

Keywords non-literal language, semantic equivalence, idiomaticity, linguistic evidence, large language models

* First author & Corresponding author

1. Introduction

In the age of AI, the relationship between linguistics and language computation can no longer be understood simply as one between a discipline and its object of study. As computational systems become capable of generating, comparing, and evaluating language at scale, their outputs raise a methodological question for linguistics: how can model behavior be interpreted as evidence about meaning interpretation rather than only as performance on predefined tasks? This question gives renewed significance to linguistic categories such as semantic equivalence, idiomaticity, and meaning interpretation. Non-literal language offers a particularly productive diagnostic site for addressing it, because expressions such as metaphors and idioms are not simply compositional strings, but forms of meaning shaped by context, convention, cultural knowledge, and varying degrees of lexicalization (Lakoff and Johnson 1980; Croft 1993; Gibbs 2008). They therefore make visible the distance between surface fluency and linguistically defined meaning: a system may produce fluent language while still failing to capture intended meaning, conventional meaning, or cross-linguistic equivalence. The Metaphor Identification Procedure (MIP) formalizes one aspect of this problem by identifying lexical items whose contextual meaning diverges from their basic sense (Pragglejaz Group 2007). Cross-linguistic idiom resources such as LIIdioms likewise show the need to represent idiomatic meaning across languages rather than reduce it to literal translation (Moussallem et al. 2018). In this sense, AI does not make linguistic analysis less necessary; it makes the evidential role of linguistics more explicit by requiring clearer definitions of what counts as meaning, understanding, and explanation.

2. AI, operationalization, and linguistic evidence

The methodological significance of AI lies in how it makes linguistic categories operational. When language models generate, compare, and evaluate language outputs at scale, concepts such as semantic equivalence, idiomaticity, and figurative interpretation can be transformed into testable conditions. They can be examined through controlled contrasts, measurable outputs, and linguistically interpretable patterns of model behavior. Prior work on idiomaticity detection and sentence

embedding has shown how such distinctions can be converted into classification and comparison tasks, while recent LLM-based evaluations demonstrate how model behavior can be used to test the boundaries of those task formulations (Madabushi et al. 2022; Phelps et al. 2024).

Prompt-based testing illustrates this shift. A prompt does more than request an answer; it specifies a relation between input, constraint, and expected output. By varying prompts, examples, answer formats, and contextual conditions, researchers can observe whether model behavior preserves a target meaning relation or moves toward more local cues. This matters for non-literal language because fluent output may still fail to recover idiomatic or figurative meaning. A prompt is therefore a theory-laden methodological operation: it encodes assumptions about the relevant cue, the expected interpretation, and the acceptable form of response (Phelps et al. 2024).

Embedding-based comparison and contrastive pair design make meaning comparison scalable. They allow researchers to move beyond exact lexical matching and examine degrees of similarity, partial equivalence, and cases in which surface form remains close while meaning diverges. Yet similarity scores remain proxies. They can reveal patterns, clusters, and distances, but they cannot decide by themselves what counts as semantic equivalence or idiomatic interpretation. The key question is therefore what linguistic judgment a score stands in for, and under what conditions that proxy is valid (Madabushi et al. 2022).

Models can also function as computational probes when their use is defined by explicit linguistic conditions. They can generate stimuli, construct contrastive settings, and make certain linguistic contrasts observable at scale. However, a probe is not an explanation by itself. Model outputs remain artifacts of a computational system and must be interpreted through linguistic standards rather than treated as direct evidence of human understanding. Work on synthetic impossible languages provides a boundary case: by testing models on artificial systems with unnatural word orders or typologically unusual rules, such studies show how model behavior can be used to examine relations among learnability, typological markedness, and linguistic possibility (Kallini et al. 2024).

Multilingual comparison increases the pressure on operational definitions. When the same task is applied across languages, it becomes easier to see whether a model is tracking meaning relations or relying on surface overlap. This matters because languages differ not only in vocabulary, but also in lexicalization, grammatical

resources, cultural semantics, and patterns of idiomatization. Cross-linguistic idiom resources and recent multilingual studies therefore show that AI creates a stringent test of whether meaning, convention, and equivalence have been defined clearly enough to travel across languages (Moussallem et al. 2018; Li et al. 2024; Khoshtab et al. 2025).

These methods do not reduce the role of linguistics; they make it more explicit. Linguists define the relevant distinctions, design the comparisons, determine what a metric can stand in for, and decide when a model pattern can be treated as evidence. Without these decisions, model outputs remain strings, scores, or rankings rather than linguistic knowledge. This methodological framework motivates the two analyses that follow. Metaphor understanding and cross-lingual idiom alignment provide complementary diagnostic perspectives on non-literal language: metaphor foregrounds projected meaning within context, whereas idiom alignment foregrounds conventionalized equivalence across languages.

3. Metaphor as a diagnostic site

Metaphor provides a useful site for examining how AI-based modeling and linguistic analysis can jointly diagnose model behavior in relation to non-literal meaning. As a subtype of figurative expression, metaphorical meaning is not derived through simple word-by-word composition. It involves a selective mapping between a source domain and a target domain, where only certain properties of the source are projected onto the target. This mapping is activated, constrained, and interpreted through context. Computational approaches to metaphor have therefore developed along several related but distinct lines: Selection with Selectional Preference Violation treats unusual or infelicitous collocations as possible signals of metaphor; the Metaphor Identification Procedure focuses on the contrast between contextual meaning and basic meaning; and Conceptual Metaphor Theory emphasizes source-target relations, selective attribute projection, and contextual interpretation (Wilks 1975; Lakoff and Johnson 1980; Croft 1993; Pragglejaz Group 2007; Gibbs 2008; Sullivan 2013). For example, in “The computer is a turtle,” the relevant meaning is not the full set of properties associated with turtles, but a projected attribute such as slowness. This makes metaphor a useful diagnostic setting; the question is not only whether a model produces a fluent

or correct answer, but what kind of semantic behavior makes that answer possible.

On this basis, we decompose LLM metaphor processing into three diagnostic dimensions: semantic attribute alignment, lexical invariance, and syntactic sensitivity. (Ye et al. 2026a). As shown in Figure 1 and Table 1, Spatial Analysis probes attribute-level semantic alignment in metaphor interpretation, Metaphorical Imagination probes the persistence of stable lexical associations under two contextual settings, and Syntactic Shuffle analyzes the influence of syntactic cues on metaphor detection. This decomposition is not intended to reduce metaphor processing to a small set of metrics. Rather, it turns the broad question of metaphor processing into three more observable questions: whether model-generated interpretations align with the semantic attribute reflected in human interpretations, whether stable lexical anchors persist across contextual and decontextualized conditions, and whether metaphor detection is affected by syntactic irregularity. In this way, metaphor processing is reformulated as a set of operational, comparable, and interpretable probes.

Across the three probes, we evaluate seven language models and use abbreviated names for consistency in tables and discussion: DeepSeek-V3-671B as V3-671B (Liu et al. 2024), Qwen2.5-Turbo as Qwen-T (Yang et al. 2024), GPT-4 (Achiam et al. 2023), GPT-4o (Hurst et al. 2024), o3-mini (OpenAI 2025), DeepSeek-R1-671B as R1-671B (Guo et al. 2025), and LLaMA-3.1-8B (Grattafiori et al. 2024). These models are used to compare metaphor-processing behavior across the three diagnostic probes, rather than to rank models by overall capability.

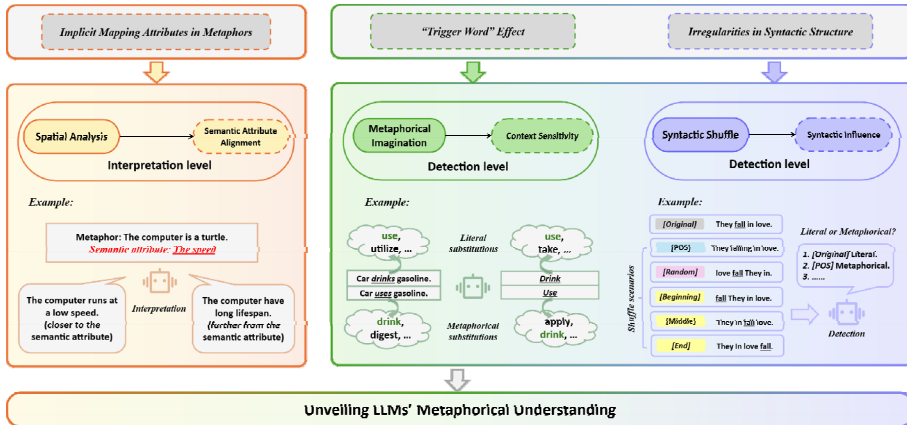


Figure 1. Overview of the experimental framework

Table 1. Dataset statistics and example instances for Fig-QA (used in spatial analysis) and MUNCH (used in metaphorical imagination and syntactic shuffle). Different MUNCH rows correspond to distinct experimental settings derived from the same set of base instances.

Examples are shown as metaphor | literal

Setting	Instances	Example (Metaphor Literal)
Fig-QA	2.6k	The computer is a race car. The computer runs fast. The computer is a turtle. The computer runs slow.
MUNCH (Context)	2.9k	The council appealed by case stated ... The council petitioned by case stated ...
MUNCH (Word)	2.9k	appealed petitioned
MUNCH (Original)	2.9k	The council appealed by case stated.
MUNCH (POS)	1.3k	The council complainant (n.) by case stated.
MUNCH (Random)	2.9k	council case appealed stated by The.
MUNCH (Beginning)	2.9k	appealed The council by case stated.
MUNCH (Middle)	2.9k	The council by case appealed stated.
MUNCH (End)	2.9k	The council by case stated appealed.

Semantic attribute alignment

Spatial Analysis targets semantic attribute alignment in metaphor interpretation. We use Fig-QA, in which each item is organized around two metaphors that share the same semantic attribute but differ in surface form, together with their corresponding human literal interpretations (Liu et al. 2022). To construct a local semantic reference region, we define a reference plane using two human interpretations and one model-generated literal paraphrase. We then define an interpretation plane using the same two human interpretations and the model-generated interpretation. The perpendicular distance d_p measures how far the model interpretation deviates from the reference plane, while $\cos \theta$ measures the directional similarity between the two planes.

Table 2. Average d_p and $\cos \theta$ across models. Mean (M) and standard deviation (SD) are reported. For d_p , lower values indicate smaller perpendicular deviation from the reference plane; for $\cos \theta$, higher values indicate stronger angular alignment. Bold and underlined values indicate the best and second-best results according to the direction of each metric; for standard deviations, lower values indicate less variability

	V3-671B	Qwen-T	GPT-4	GPT-4o	o3-mini	R1-671B	LLaMA-3.1-8B
d_{pM}	0.1903	0.2319	0.2267	0.1772	0.2020	0.2063	0.2866
$(\cos \theta)_M$	0.8207	0.7835	0.7940	0.7526	0.7931	0.7804	0.7396
d_{pSD}	0.2194	0.2386	0.2342	0.2182	0.2343	0.2204	0.2649
$(\cos \theta)_{SD}$	0.2531	0.2742	0.2669	0.2905	0.2703	0.2698	0.2918

A small-scale human validation supports this interpretation. Low- d_p interpretations receive substantially higher human alignment scores than high- d_p interpretations, with a mean contrast of 1.96 versus 0.84 and a statistically significant difference ($p \approx 1.05 \times 10^{-6}$). This does not make d_p an absolute measure of semantic alignment, but it suggests that geometric deviation tracks a meaningful semantic difference. The broader point is that metaphor interpretation cannot be evaluated by fluency alone: a model may move toward the right semantic region without consistently aligning with human attribute-based interpretation.

Lexical invariance under contextual variation

Metaphorical Imagination examines lexical invariance in metaphor processing: the extent to which models generate similar substitute expressions under contextualized and decontextualized conditions. This responds to what can be described as a trigger-word effect: a model may be guided by highly associated lexical items rather than by the metaphorical mapping licensed by a specific context (Wachowiak and Gromann 2023). We use MUNCH and compare two generation directions: literal-to-metaphor and metaphor-to-literal. In each direction, the model generates substitute words either from the full sentence context or the target metaphorical word alone (Tong et al. 2024).

To quantify this behavior, we use an Anchor Score. If the contextualized and decontextualized output sets share words, the score is set to 1. If they do not share words, the score is based on the maximum pairwise cosine similarity between the two output sets in a 300-dimensional GloVe space (Pennington et al. 2014). This measure captures both lexical overlap and near-overlap as behavioral signals, making

it more suitable for this substitution setting than strict string matching. Table 3 reports Anchor Scores across models, generation directions, and discourse genres.

Table 3. Anchor scores (%) for metaphorical imagination under literal-to-metaphor (LM) and metaphor-to-literal (ML) settings. Genre-level results are reported for News, Fiction, Academic, and Conversation. Highest values within each row are shown in bold, and second-highest values are underlined

	V3-671 B	Qwen- T	GPT- 4	GPT-4 o	o3-mi ni	R1-671 B	LLaMA-3.1-8 B
LM	73.30	73.45	78.92	76.28	72.25	78.11	65.09
ML	76.96	75.17	79.22	78.01	81.55	80.87	72.86
News (LM)	74.43	74.91	80.96	76.90	72.76	78.00	66.28
News (ML)	77.82	75.23	83.77	79.32	81.27	80.68	72.11
Fiction (LM)	73.95	73.55	76.06	79.30	68.20	76.22	64.32
Fiction (ML)	75.21	73.19	74.73	75.70	78.73	77.97	67.79
Academic (LM)	75.02	76.11	82.71	77.15	71.09	80.37	65.46
Academic (ML)	80.69	79.69	84.93	81.84	85.25	83.62	75.09
Conversation (LM)	66.80	66.17	69.74	71.20	73.22	75.00	61.27
Conversation (ML)	73.35	71.13	71.16	73.86	81.87	81.11	73.94

The results in Table 3 show substantial lexical anchor persistence across directions and contextual settings. Anchor Scores are consistently high, generally falling in the 65–85% range, with metaphor-to-literal generation usually producing higher scores than literal-to-metaphor generation. This suggests that model behavior remains strongly shaped by stable lexical associations, even when sentence context is available. Such stability is not necessarily a failure. For conventional metaphors, lexicalized associations may support plausible interpretation. However, in more novel or context-sensitive metaphors, the same stability may bias the model toward familiar collocations or high-frequency associations, thereby limiting contextual integration.

This finding should be interpreted cautiously. A high Anchor Score does not prove that context is ignored, because lexical priors and contextual evidence may point in the same direction. The narrower conclusion is that stable lexical associations remain available across conditions and can shape model behavior. In this sense, observed metaphor-processing behavior may reflect a mixture of lexical association, statistical co-occurrence, and contextual integration rather than a single attribute-level semantic

mechanism.

Syntactic influence and structural perturbation

Syntactic Shuffle examines whether metaphor detection is affected by syntactic structure and word-order irregularity. Using MUNCH, we apply controlled perturbations while preserving part of the lexical content. Random Shuffle disrupts overall word order and sentence coherence; POS Shuffle replaces the metaphorical word with a near-synonymous alternative of a different part of speech, introducing local syntactic irregularity; and Metaphorical Word Reposition moves the metaphorical word to the beginning, middle, or end of the sentence. The aim is to test whether models integrate syntactic and semantic structure, or whether detection behavior is shaped by local anomaly, lexical salience, or positional cues (Sullivan 2013; Tong et al. 2024).

Table 4. Metaphor detection accuracy under syntactic shuffle perturbations. Original, Random, and POS report the original and structural perturbation conditions, while Beginning, Middle, and End report positional perturbations of the metaphorical word. Bold values indicate the highest accuracy within each row

	Original	Random	POS	Beginning	Middle	End
V3-671B	18.31	22.71	23.59	19.14	22.10	22.63
Qwen-T	30.37	1.17	33.59	25.46	27.15	27.50
GPT-4	34.73	12.93	43.74	36.07	37.92	37.60
GPT-4o	28.89	7.78	36.87	30.92	30.84	29.98
o3-mini	29.87	5.40	38.63	25.27	25.85	25.58
R1-671B	28.68	12.22	46.41	39.25	30.88	36.03
LLaMA-3.1-8B	53.36	50.33	53.81	51.75	53.08	53.67

As shown in Table 4, syntactic perturbation substantially changes metaphor detection behavior. In several models, POS Shuffle produces higher accuracy than the original sentence. For example, R1-671B rises from 28.68 in the original condition to 46.41 under POS Shuffle, while Qwen-T rises from 30.37 to 33.59. This does not mean that the perturbation improves metaphor detection ability. A more cautious interpretation is that syntactic irregularity itself may function as a cue for metaphoricality. In such cases, the model may treat metaphoricality partly as anomaly rather than as meaning transfer alone.

Random Shuffle requires separate caution because it disrupts both structure and

coherence, making it better understood as a stress test than as a naturalistic condition. Positional perturbations are generally weaker than POS Shuffle, suggesting that models are more responsive to local structural irregularity than to word position alone. LLaMA-3.1-8B provides a useful low-responsiveness case: it remains around the 50% range and varies only modestly across perturbation types. Taken together, these results suggest that metaphor-related behavior in current LLMs is sensitive to semantic attribute alignment, lexical anchoring, and syntactic sensitivity, rather than being governed by a single stable mechanism.

Together, the three probes show why metaphor is useful as a diagnostic site. Spatial Analysis asks whether generated interpretations approach the expected semantic attribute region; Metaphorical Imagination tests whether lexical anchors persist across contextual conditions; and Syntactic Shuffle examines whether detection behavior is influenced by structural irregularity. These signals separate behavioral patterns that may otherwise be collapsed under task accuracy and make model behavior more linguistically interpretable.

4. Cross-lingual idiom alignment as a diagnostic site

Cross-lingual idiom alignment provides a complementary site for examining how AI-based modeling and linguistic analysis can jointly diagnose model behavior in relation to non-literal meaning across languages. Unlike metaphor, which foregrounds meaning construction within a language, idiom alignment foregrounds meaning equivalence between languages. Idiomatic meanings are often conventionalized and non-compositional, making literal transfer unreliable (He et al. 2025). In cross-lingual settings, the crucial question is not simply whether a model recognizes an idiom, but whether it can identify a target-language idiom with the same conventionalized figurative meaning (Moussallem et al. 2018; Donthi et al. 2025). This makes idiom alignment a diagnostic setting for separating literal transfer, lexical association, contextual plausibility, semantic paraphrase, and idiom-to-idiom equivalence, especially because recent work shows that LLMs may still rely on surface or contextual cues that do not guarantee idiomatic equivalence (Mi et al. 2025; Yang et al. 2025).

Existing idiom resources and benchmarks have operationalized idiomaticity from several perspectives. Detection and disambiguation datasets separate literal from

figurative readings (Haagsma et al. 2020; Tedeschi et al. 2022; De Luca Fornaciari et al. 2024), while contextualized-embedding and contrastive studies examine whether idiomaticity is encoded in model representations or can be detected as non-compositionality (Fakharian and Cook 2021; Zhou et al. 2023). Generation, paraphrase, multilingual, and translation-oriented benchmarks further extend idiom evaluation across tasks and languages (Zhou et al. 2021; Khoshtab et al. 2025). These lines of work are valuable, but they do not directly isolate the controlled question at the center of idiom alignment: whether two expressions in different languages should be treated as meaning-equivalent idioms rather than literal equivalents, plausible paraphrases, or merely related phrases.

4.1 Gloss-pivoted benchmark construction

On this basis, G-IdiomAlign uses English glosses as a semantic pivot for constructing and evaluating cross-lingual idiom equivalence (Ye et al. 2026b). The gloss is not treated as the final answer. Rather, it functions as a controlled meaning description that helps reduce reliance on surface overlap. Idiom entries and core glosses are extracted from Wiktionary, usage notes and examples are excluded, and single-sense idioms are retained to avoid ambiguous one-to-many sense correspondences. The construction pipeline then retrieves top- K candidates in a shared gloss-embedding space, applies mutual nearest-neighbor alignment, and uses language-pair-specific distribution-aware filtering to retain high-confidence pairs. The resulting core benchmark contains 18,785 aligned idiom pairs across 36 unordered language pairs drawn from nine core languages: German, English, Spanish, Finnish, French, Japanese, Polish, Chinese, and Portuguese. Since each unordered pair can be evaluated in both directions, the open-ended generation setting covers 72 ordered directions. Figure 3 summarizes the construction pipeline and Table 5 provides representative examples of aligned idiom pairs and their glosses.

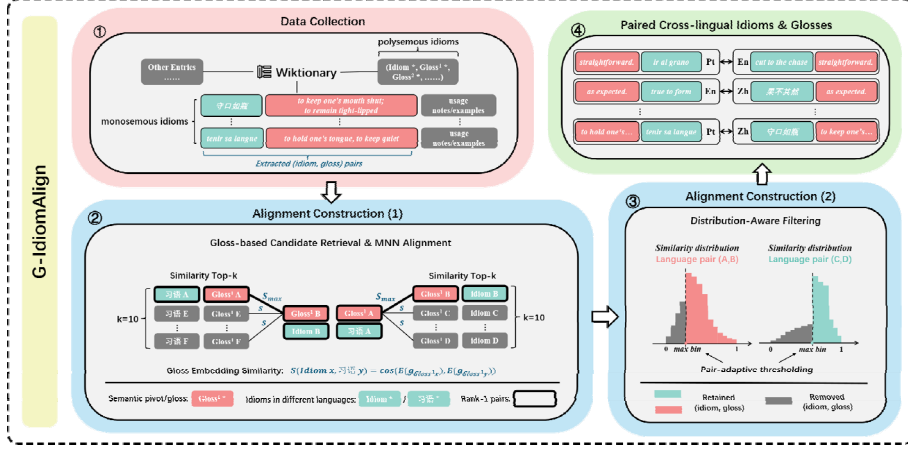


Figure 3. Overview of the G-IdiomAlign construction pipeline

Table 5. Example idiom pairs from G-IdiomAlign across different language pairs. I and G denote idioms and their English glosses; subscripts indicate source or target languages

I_{source}	I_{target}	G_{source}	G_{target}
一鋪清袋	lose one's shirt	lose all money at once	lose all of the money
mouton de Panurge	follow ... off a cliff	blindly follow others	follow a leader blindly
守口如瓶	tenir sa langue	keep one's mouth shut	hold one's tongue

The reliability of the reference set is assessed with a three-point semantic equivalence scale: 2 for full equivalence, 1 for partial equivalence, and 0 for non-equivalence. Strict accuracy counts only score-2 pairs, while lenient accuracy counts both score-1 and score-2 pairs. For non-Zh-En language pairs, LLM-based majority-vote validation yields a mean strict accuracy of 0.685 and a mean lenient accuracy of 0.923 when each language pair is treated as one observation. For Zh-En, human evaluation over 200 sampled pairs yields 0.655 strict accuracy and 0.895 lenient accuracy. These numbers indicate that G-IdiomAlign is designed as a high-confidence diagnostic resource rather than a loose set of semantically associated idioms. Building on this resource, the following evaluation compares seven model settings: DeepSeek-V3.2 (NT/T), Gemini-2.5-Pro, Claude-4.5-Haiku, Ministral-8B-Instruct, and Qwen3-8B (NT/T) (Mistral AI 2024; Anthropic 2025; DeepSeek-AI et al. 2025; Google

Gemini Team 2025; Yang et al. 2025). Here, NT and T denote No-thinking and Thinking modes, respectively.

4.2 Multiple-choice idiom equivalence

The evaluation decomposes cross-lingual idiom alignment into two observable probes. The first, Multiple-Choice Idiom Equivalence, turns the alignment problem into a controlled four-way decision with typed distractors. Each item contains one canonical target idiom and three error-inducing alternatives: a Literal Translation Trap (LT), a Lexical Cue Trap (LC), and a Contextual Association Trap (CA). These distractors correspond to different ways in which a model may confuse meaning levels: it may follow word-by-word transfer, rely on partial lexical overlap, or choose an expression that is contextually plausible but not semantically equivalent. The task covers 30 ordered direction settings, grouped by Chinese-target (Zh-target), English-target (En-target), and other target languages. To ensure that the distractors do not introduce superficial shortcuts, the benchmark includes an option-only control and a manual verification of 200 questions, with 81.5% distractor-type validity. Table 6 reports multiple-choice accuracy aggregated by target-language group.

Table 6. Multiple-choice accuracy aggregated by target-language groups. Micro is instance-weighted within each target group (Zh-/En-/Other target), while macro averages over directions. All numbers are percentages

Model	Zh-target		En-target		Other targets		Overall	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
DeepSeek-V3.2 (NT)	56.84	56.96	57.84	56.16	47.39	47.53	52.70	53.65
DeepSeek-V3.2 (T)	69.37	69.18	66.70	65.38	53.92	54.09	61.45	63.02
Gemini-2.5-Pro	67.77	67.57	66.53	64.95	62.82	62.91	65.13	65.17
Claude-4.5-Haiku	56.80	57.30	55.54	53.31	44.03	43.44	50.50	51.47
Ministral-8B-Instruct	28.63	28.51	30.29	28.97	27.78	27.76	28.68	28.43
Qwen3-8B (NT)	32.69	33.72	33.95	33.62	24.02	23.99	28.98	30.56
Qwen3-8B (T)	41.20	41.89	39.57	39.10	31.29	31.39	36.14	37.55

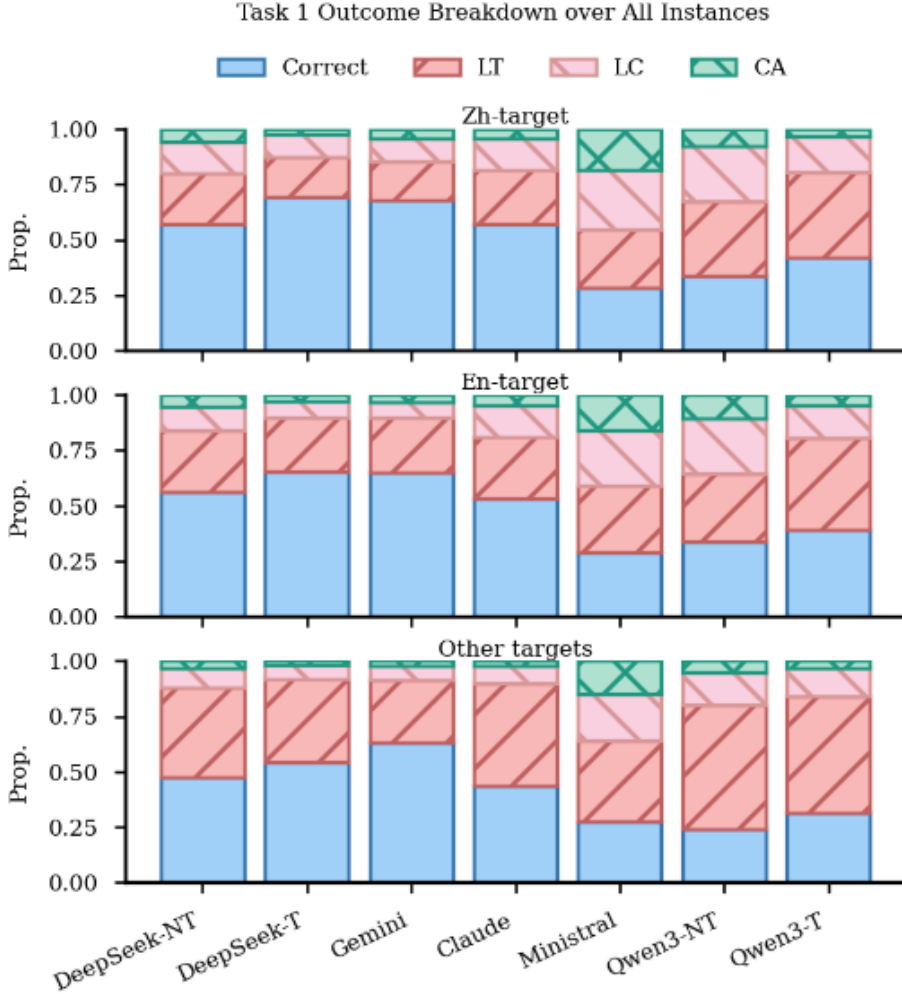


Figure 4. Task 1 outcomes by target-language regime. Stacked bars (normalized within each regime) decompose model responses into correct predictions and LT/LC/CA, highlighting differences between high-resource targets (Zh/En) and other target languages

As shown in Table 6, while the multiple-choice probe does more than measure overall answer accuracy, it also makes visible which linguistic cue attracts the model away from idiomatic equivalence. Across models, accuracy is lower when the target language belongs to the “Other targets” group than when the target is Chinese or English, suggesting that target-language resource conditions affect idiom alignment.

Figure 4 further shows the distribution of distractor choices: literal-translation attraction dominates the error cases, while Lexical Cue errors reveal shallow lexical matching and Contextual Association errors show that contextual relevance does not guarantee semantic equivalence. Thus, the task reformulates cross-lingual alignment errors as a linguistic diagnostic question: which level of meaning has the model followed?

4.3 Gloss-contrastive generation

The second probe, Gloss-Contrastive Generation, moves from controlled selection to open-ended production. The model must generate a meaning-equivalent idiom directly in the target language. The task contrasts two input conditions: No-gloss, where the model receives only the source idiom, and With-gloss, where it also receives the English gloss. This contrast tests whether an explicit semantic cue helps the model move from source-language form toward target-language idiomatic meaning. Because open-ended idiom generation permits multiple possible outputs, evaluation uses an embedding-based semantic proxy: model outputs and canonical target idioms are encoded with Qwen3-Embedding-8B (Zhang et al. 2025), and performance is reported with MeanSim and $\text{Acc}@t$ at $t = 0.70$ and $t = 0.80$. These measures are useful for comparing trends across all 72 directions, but they are not treated as definitive correctness judgments, since a valid non-canonical idiom may be missed and a semantically close paraphrase may still fail to be idiomatic.

Table 7. Gloss-contrastive generation results. MeanSim ($100 \times \text{sim}$) and $\text{Acc}@t$ are direction-averaged (macro); $\text{Acc}@t$ counts a prediction as correct if cosine similarity $\geq t$. We report $t = 0.70$ and 0.80 . $\Delta\text{Acc}@t$ is the With-gloss minus No-gloss improvement. Bold indicates the best performance

Model	No-gloss			With-gloss			
	MeanSim	Acc@0.7	Acc@0.8	MeanSim	Acc@0.7	Acc@0.8	$\Delta\text{Acc}@0.8$
DeepSeek-V3.2 (NT)	68.30	41.99	19.90	70.94	48.35	23.57	3.66
DeepSeek-V3.2 (T)	72.31	49.01	24.66	74.24	54.90	29.63	4.97
Gemini-2.5-Pro	68.71	44.06	19.70	72.54	51.09	27.11	7.41
Claude-4.5-Haiku	70.90	45.47	20.47	72.94	52.16	25.66	5.19

	No-gloss			With-gloss			
Ministral-8B-Instruct	68.30	37.97	12.96	71.69	50.00	21.22	8.28
Qwen3-8B (NT)	67.54	36.29	11.59	70.69	47.09	18.80	7.22
Qwen3-8B (T)	68.53	39.03	14.04	71.40	49.16	20.86	6.82

As shown in Table 7, With-gloss improves performance over No-gloss across all reported models. These results show that explicit semantic support helps, but the improvement remains limited. Glosses can guide the model toward the intended meaning, yet they do not automatically produce canonical target-language idioms. In diagnostic terms, the gloss helps with semantic anchoring, but it does not by itself solve the transition from meaning recovery to idiomatic form.

4.4 Attention-based diagnostics

Attention-based diagnostics provide a further, explicitly correlational view of this process. Using Qwen3-8B under greedy decoding, the analysis extracts post-softmax self-attention and compares No-gloss and With-gloss generations. Both conditions contain 189 valid annotated cases; two annotators label the outputs with Cohen’s $\kappa = 0.81$. The diagnostics define IAR as attention mass on the idiom span, GAR as attention mass on the gloss span, DR as residual attention mass outside the tracked spans, and OtherTop1 as the strongest off-span concentration. These measures serve as observable, correlational signals of where the model allocates attention while producing correct or incorrect outputs.

Table 8. Cross-condition overlap between With-gloss and No-gloss for salient head sets and salient layer sets

Subset	J_{heads}	J_{layers}
All	0.32	0.90
Correct-only	0.32	0.76
Wrong-only	0.37	0.90

Table 9. Token-level attention diagnostics (correct vs. Wrong). Each metric measures where the model attends during generation: IAR, GAR, DR, and OtherTop1. We report median (IQR) for each group, Cliff’s δ and q -values from two-sided Mann-Whitney U tests with BH-FDR correction within each condition

Metric	Analysis	With-gloss				No-gloss			
		Correct	Wrong	δ	q	Correct	Wrong	δ	q
IAR	CvsW	0.03 (0.01)	0.03 (0.01)	-0.1 0	0.25 0	0.06 (0.02)	0.06 (0.02)	0.1 8	0.08 7
GAR	CvsW	0.07 (0.04)	0.06 (0.03)	0.2 4	0.02 6	--	--	--	--
DR	CvsW	0.90 (0.04)	0.91 (0.03)	-0.2 0	0.03 3	0.94 (0.02)	0.94 (0.02)	-0.1 8	0.08 7
OtherTop1	CvsW	0.57 (0.03)	0.58 (0.02)	-0.2 1	0.03 3	0.58 (0.02)	0.58 (0.02)	0.0 1	0.90 0

The structural comparison in Table 8 shows that cross-condition differences are concentrated more in attention heads than in layers. Jaccard similarity measures the overlap between the salient sets in the No-gloss and With-gloss conditions; overall head overlap is much lower than layer overlap, with $J_{\text{heads}} = 0.32$ and $J_{\text{layers}} = 0.90$. This suggests that the two conditions largely recruit similar layers but differ in which heads within those layers become salient. The token-level diagnostics in Table 9 further show that correct With-gloss generations exhibit stronger gloss anchoring than wrong With-gloss generations: Comparing correct and wrong scenarios, the GAR result is higher, while DR and OtherTop1 are lower. These patterns indicate that successful With-gloss generation is associated with stronger use of the gloss as a semantic anchor, although this anchoring remains insufficient for fully reliable idiom generation.

In this work, cross-lingual idiom alignment reformulates idiom-processing behavior as a set of observable diagnostic contrasts. Multiple-Choice Idiom Equivalence distinguishes idiomatic equivalence from literal transfer, lexical cueing, and contextual association. Gloss-Contrastive Generation tests whether an explicit meaning cue helps models move from source-language form to target-language idiomatic meaning. Attention-based diagnostics further ask whether successful generation is associated with stronger semantic anchoring. In this way, we treat idiom alignment functions not simply as a translation benchmark, but as a diagnostic site for identifying how model behavior moves among literal meaning, semantic paraphrase, and idiom-to-idiom equivalence across languages.

5. Beyond task performance: Linguistic interpretability and evidence

The preceding analyses show that model behavior in non-literal language should not be interpreted only as task performance. Accuracy, similarity scores, generation quality, and attention patterns are useful starting points, but their linguistic value depends on what kind of meaning relation they are taken to index. In the metaphor probes, improved or reduced performance may reflect different combinations of semantic attribute alignment, lexical anchoring, and syntactic sensitivity. In the idiom probes, model behavior may reflect literal transfer, lexical cueing, contextual plausibility, gloss-based semantic anchoring, or idiom-to-idiom equivalence. The central question is therefore not simply whether task performance is high or low, but which linguistically defined relation a model appears to follow in a given setting.

This perspective reframes both successful and unsuccessful outputs as interpretable behavioral evidence. A fluent or correct output may still depend on a local cue rather than the intended relation, while an incorrect output may reveal which cue has become salient in place of that relation. Conversely, performance improvement may show that an added condition, such as the addition of a gloss or a syntactic perturbation, changes which information the model can use. Such patterns should not be read as direct evidence of human-like understanding. They become linguistically informative only when they are connected to explicit contrasts such as literal versus figurative meaning, paraphrase versus equivalence, context versus convention, and surface cue versus meaning relation.

This is why interpretability cannot be treated as a purely technical problem. Attention patterns, embedding distances, similarity scores, and generated outputs are computational traces, not explanations by themselves. Their significance depends on the linguistic categories through which they are interpreted. Linguists therefore contribute not only by evaluating outputs after the fact, but by defining the categories, contrasts, datasets, and evidential standards that make model behavior interpretable. In this sense, AI does not replace linguistic analysis; it increases the need for linguistic analysis to structure what counts as evidence.

The relationship between linguistics and AI is therefore best understood as a shared methodological space. AI methods make it possible to generate stimuli, compare outputs, and observe behavior across models, prompts, languages, and task formats at scale. Linguistic analysis determines what those observations mean. The metaphor

and idiom analyses in this article illustrate this division of roles: computational procedures make meaning-related behavior observable, while linguistic categories make it interpretable as evidence about language rather than merely as performance on a task.

6. Conclusion

This article has examined non-literal language as a diagnostic site for rethinking linguistic evidence in the age of AI. Non-literal language is especially suited to this purpose since its literal form does not straightforwardly convey the intended meaning. It therefore forces a distinction between surface form and target meaning: a fluent output may still miss a projected semantic attribute, rely on a lexical cue, or approximate a meaning without recovering a conventionalized idiomatic equivalent. Accuracy, similarity scores, generation quality, and attention patterns become linguistically meaningful only when they are interpreted through such explicit categories of meaning.

The two presented analyses make this argument concrete. The metaphor probes distinguish semantic attribute alignment, lexical anchoring, and syntactic sensitivity, while the idiom probes distinguish literal transfer, lexical cueing, contextual plausibility, gloss-based semantic anchoring, and idiom-to-idiom equivalence. These contrasts show that model behavior cannot be fully understood by asking whether performance is high or low. The more important question is which meaning relation a model appears to follow in a given setting.

The broader implication is that AI does not reduce the role of linguistic analysis; it makes that role more explicit. LLMs provide a new empirical setting in which meaning-related behavior can be generated, compared, and observed at scale. Linguistics provides the categories, contrasts, and evidential standards needed to interpret that behavior. In this sense, the relationship between AI and linguistics is not one of replacement, but of methodological interaction: computational procedures make model behavior observable, while linguistic analysis makes it interpretable as evidence about language.

References

- Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2025. Introducing Claude Haiku 4.5. Anthropic.
- Croft, William. 1993. The role of domains in the interpretation of metaphors and metonymies. *Cognitive Linguistics* 4(4): 335–370.
- De Luca Fornaciari, Francesca, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty, and Emmy Liu (eds.), *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, 35–44. Mexico City: Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Donthi, Sundesh, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage (eds.), *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, 175–181. Abu Dhabi: Association for Computational Linguistics.
- Fakharian, Samin and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, 23–32. Online.
- Gibbs, Raymond W. Jr. (ed.). 2008. *The Cambridge handbook of metaphor and thought*. Cambridge: Cambridge University Press.
- Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multi-modality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Haagsma, Hessel, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation*

- Conference*, 279–287. Marseille: European Language Resources Association.
- He, Wei, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics* 51(2): 505–555.
- Hurst, Aaron, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- Kallini, Julie, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14691–14714. Bangkok: Association for Computational Linguistics.
- Khoshtab, Paria, 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*, 8680–8698. Abu Dhabi: Association for Computational Linguistics.
- Lakoff, George and Mark Johnson. 1980. *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Li, Shuang, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: IdiomKB’s role in optimizing idiomatic translation with language models. In Michael Wooldridge, Jennifer Dy, and Sriraam Natarajan (eds.), *Proceedings of the AAAI Conference on Artificial Intelligence* 38(17), 18554–18563. Washington, DC: AAAI Press.
- Liu, Aixin, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.
- Liu, Emmy, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4437–4452. Seattle, WA: Association for Computational Linguistics.
- Mi, Maggie, 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)*, 7314–7332. Vienna: Association for Computational Linguistics.
- Mistral AI. 2024. Un ministral, des ministraux. Mistral AI.
- Moussallem, Diego, 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: European Language Resources Association.

- OpenAI. 2025. OpenAI o3-mini system card. OpenAI.
- Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha: Association for Computational Linguistics.
- Phelps, Dylan, 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*, 178–187. Torino: ELRA and ICCL.
- Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22(1): 1–39.
- Sullivan, Karen. 2013. *Frames and constructions in metaphoric language*. Amsterdam: John Benjamins Publishing Company.
- Tayyar Madabushi, Harish, 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)*, 107–121. Seattle, WA: Association for Computational Linguistics.
- Tedeschi, Simone, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2715–2726. Seattle, WA: Association for Computational Linguistics.
- Tong, Xiaoyu, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3517–3536. Bangkok: Association for Computational Linguistics.
- Wachowiak, Lennart and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1018–1032. Toronto: Association for Computational Linguistics.
- Wilks, Yorick A. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence* 6(1): 53–74.
- Yang, An, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yang, An, 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, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Yang, Cai, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025. Evaluating LLMs

- on Chinese idiom translation. *arXiv preprint arXiv:2508.10421*.
- Ye, Fengying, Shanshan Wang, Lidia S. Chao, and Derek F. Wong. 2026a. Probing semantic alignment, lexical invariance, and syntactic influence in LLM metaphor processing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 27914–27932. San Diego, California, United States: Association for Computational Linguistics.
- Ye, Fengying, Yanming Sun, Runzhe Zhan, Lidia S. Chao, Zheqi Zhang, and Derek F. Wong. 2026b. G-IdiomAlign: A gloss-pivoted benchmark for cross-lingual idiom alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 38720–38739. San Diego, California, United States: Association for Computational Linguistics.
- Zhang, Yanzhao, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Zhou, Jianing, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, 33–48. Online.
- Zhou, Jianing, Ziheng Zeng, and Suma Bhat. 2023. CLCL: Non-compositional expression detection with contrastive learning and curriculum learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 730–743. Toronto: Association for Computational Linguistics.

Derek F. Wong

Professor
Department of Artificial Intelligence
University of Macau
Avenida da Universidade
Taipa, Macau, China
E-mail: derekfw@um.edu.mo

Fengying Ye

PhD Candidate
Department of Artificial Intelligence
University of Macau
Avenida da Universidade

Taipa, Macau, China

E-mail: nlp2ct.fengying@gmail.com

Received: 2026. 06. 21.

Revised: 2026. 07. 30.

Accepted: 2026. 08. 03.