VIVID: A Culturally Grounded Benchmark Exposing the Figurative Language Gap in Vietnamese NLP

Tu Tran Do[1*], **Nhat Ngoc Nguyen**[1*], **Khanh-Tung Tran**[2]
Hoang D. Nguyen[2], **Tu Minh Phuong**[1], **Long Hoang Dang**[1]
[1]Posts and Telecommunications Institute of Technology (PTIT), Vietnam
[2]School of Computer Science and Information Technology, University College Cork, Ireland
{tudt.b22kh109, nhatnn.b22at208}@stu.ptit.edu.vn
123128577@umail.ucc.ie hn@cs.ucc.ie
{phuongtm, longdh}@ptit.edu.vn

Abstract

We present VIVID (Vietnamese Idioms for Validation and Interpretation Depth), the first systematic benchmark for evaluating culturally grounded figurative language understanding in Vietnamese. VIVID comprises 1,636 idioms and proverbs annotated with five complexity traits (literal expressions, pragmatic nuances, Sino-Vietnamese terms, uncommon vocabulary, folk knowledge) and seven semantic themes. We establish an evaluation framework combining generative and discriminative tasks, proposing an LLM-as-a-Judge approach with aspect-based prompting validated against human judgment (Cohen's κ = 0.792). Evaluating eight state-of-the-art models reveals critical gaps: Vietnamese-specialized models drastically underperform multilingual systems (VinaLLaMA-7B: 0.13 vs. GPT-4o: 2.46), and even top models achieve less than 50% correctness on average. Notably, few-shot prompting does not universally improve performance, with GPT-4o exhibiting degradation due to stylistic overfitting. Our analysis exposes systematic failures including literal over-interpretation, lexical gaps, and pragmatic flattening, demonstrating that current models lack cultural competence for nuanced figurative interpretation. VIVID provides an essential tool for advancing figurative language understanding in culturally rich contexts. We release codes and datasets at https://github.com/ReML-AI/VIVID.

Keywords: Vietnamese Benchmark, Figurative Language Evaluation, Idioms and Proverbs

1. Introduction

Figurative language understanding remains a critical challenge in natural language processing, requiring models to integrate linguistic knowledge, cultural context, and pragmatic reasoning (Liu et al., 2022a). In Vietnamese linguistics, an idiom is defined as a fixed multi-word expression whose meaning cannot be straightforwardly inferred from the literal meanings of its constituent words. A proverb, in contrast, is a short, often rhythmic sentence that encapsulates folk knowledge, life experience, or moral lessons (Phê, 2021). Vietnamese idioms and proverbs present unique interpretation difficulties through archaic Sino-Vietnamese vocabulary, subtle pragmatic nuances (sarcasm, irony), and folk knowledge about traditional agricultural practices. Despite Vietnamese having substantial NLP resources, including benchmarks for natural language inference (Huynh et al., 2022a), sentiment analysis (Ho et al., 2019a), and multi-hop reasoning (Le et al., 2022a), **no systematic evaluation framework exists for assessing culturally grounded figurative language understanding**.

This gap motivates two research questions. **First, how do state-of-the-art language models perform on Vietnamese figurative language?** Existing benchmarks focus on high-resource languages like English (Chakrabarty et al., 2022), Korean (Wang et al., 2024), or Ethiopian languages (Azime et al., 2025) evaluating binary idiomaticity detection rather than cultural understanding. Without a Vietnamese benchmark, we cannot assess whether Vietnamese-specialized or multilingual models accurately interpret idioms, nor identify which linguistic features challenge these systems. **Second, how can we reliably evaluate figurative language understanding?** Traditional metrics like BLEU fail to capture semantic nuances (Tang et al., 2024a; Wang et al., 2025a), while human evaluation is expensive and unscalable.

To address these questions, we introduce **VIVID** (Vietnamese Idioms for Validation and Interpretation Depth), a novel benchmark of 1,636 Vietnamese idioms and proverbs with dual-layer annotations. First, we develop a linguistically-motivated taxonomy organizing expressions into three aspects (Meaning, Lexical, Agricultural & Customary Knowledge) with five complexity traits that capture primary interpretation challenges. Second, we categorize idioms into seven semantic themes (Criticism, Life Lessons, Work and Nature, Social Relationships, Love, Virtues, Other) reflecting Vietnamese cultural discourse. We evaluate eight state-of-the-art models—including Vietnamese-

[*] These authors contributed equally.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5414–5430
11–16 May 2026.

specialized systems (Vistral-7B, VinaLLaMA-7B, GreenMind-14B) and large multilingual models (GPT-4o, Gemini Flash 2.5, Llama-4-Scout, Qwen-3,Llama-SEA-LION-v3)—across complementary evaluation tasks.

For evaluation, we develop a comprehensive evaluation framework combining generative and discriminative approaches. For generative evaluation, we utilize an LLM-as-a-Judge framework where models generate idiom explanations evaluated by GPT-4.1. Through systematic comparison of four prompting strategies, we establish that aspect-based evaluation—explicitly considering factual accuracy, completeness, and native speaker alignment—achieves strongest human correlation (Cohen's $\kappa = 0.792$). For discriminative evaluation, we apply multiple-choice classification tasks for topic and linguistic characteristic identification using exact-match evaluation (Gao et al., 2024).

Our findings reveal critical gaps across all model types: Even state-of-the-art models achieve less than 50% correctness on average, with significant variation across complexity traits—models handle Sino-Vietnamese terms better than pragmatic nuances or folk knowledge. Among comparable-sized models, Vietnamese-specialized systems (GreenMind-14B: 1.09) perform similarly to multilingual counterparts (Qwen-3-14B: 1.09), suggesting that current language-specific training does not provide advantages for cultural figurative understanding. Notably, few-shot prompting does not universally help: while benefiting most models, GPT-4o shows degradation due to stylistic overfitting.

Our contributions are threefold: **(1)** VIVID, the first systematic benchmark for Vietnamese figurative language with 1,636 validated idiom-explanation pairs and dual-layer annotations; **(2)** evaluation protocols combining a novel LLM-as-a-Judge framework for generative assessment with discriminative classification tasks; and **(3)** comprehensive analysis showing model scale matters more than language-specific training, while exposing persistent cultural understanding challenges. VIVID provides a foundation for advancing figurative language understanding and developing culturally aware NLP systems.

2. Related Work

2.1. Idiom and Proverb Datasets

Existing figurative language datasets focus primarily on high-resource languages. (Chakrabarty et al., 2022) introduce FLUTE, a synthetic dataset of 1000 English idiomatic sentences paired with synonymous or contradictory statements. (Tayyar Madabushi et al., 2022) present SemEval-2022 Task 2 for multilingual idiomaticity detection across 8000 sentences in English, Portuguese, and Galician. (Azime et al., 2025) compile ProverbEval with cultural proverbs from four Ethiopian languages and English, while (Wang et al., 2024) present KorID, a cloze-test dataset of 1631 Korean idioms. These datasets advance figurative language resources but lack coverage of Vietnamese, leaving a gap in evaluating culturally grounded expressions in this language.

2.2. Vietnamese Benchmarks

Vietnamese NLP has established benchmarks for natural language inference (Huynh et al., 2022a), sentiment analysis (Ho et al., 2019a), hate speech detection (Hoang et al., 2023), and comprehensive NLU evaluation through ViGLUE (Tran et al., 2024) and VLUE (Tran et al., 2024). VIMQA (Le et al., 2022b) evaluates multi-hop reasoning capabilities. However, no dataset exists for assessing figurative and cultural language understanding in Vietnamese idioms and proverbs. VIVID addresses this gap as the first benchmark for evaluating culturally grounded figurative expressions in Vietnamese.

2.3. Figurative Language Evaluation

Evaluating figurative language understanding presents unique challenges, as traditional automatic metrics like BLEU and ROUGE fail to capture semantic nuances in idiomatic interpretation (Tang et al., 2024a; Wang et al., 2025a). Recent studies demonstrate that LLM-as-a-Judge approaches (Zheng et al., 2023) achieve higher correlation with human judgment for figurative language tasks. (S. Rezaeimanesh, 2025) shows strong alignment for Persian-English idiom translation, while (Tang et al., 2024a) and (Wang et al., 2025b),validate similar effectiveness for East Asian and Chinese idioms. However, existing work primarily establishes that LLM judges outperform automatic metrics without systematically investigating how prompting strategies affect evaluation reliability for figurative language. Different approaches—zero-shot scoring, demonstration examples, chain-of-thought reasoning, or aspect-based evaluation—may yield varying alignment with human judgment, yet this remains underexplored for culturally grounded expressions. We address this gap through systematic comparison of four prompting strategies for Vietnamese idiom evaluation, establishing that aspect-based evaluation achieves strongest human alignment (Cohen's $\kappa = 0.792$). This provides methodological guidance for reliable figurative language assessment in low-resource contexts.

3. Dataset

We present VIVID, a culturally grounded evaluation benchmark dataset contains 1636 idioms and proverbs across 7 topics that capture Vietnamese cultural contexts, ranging from everyday life to specific subject areas, as well as Vietnamese grammar and linguistics

We further annotate the dataset into five distinct characteristics that LLMs are highly prone to misunderstanding. Because idioms may exhibit multiple characteristics simultaneously, the total frequency counts exceed the number of idioms. Dataset statistics are reported in Figure 2 and 3. The dataset is constructed in three stages: 3.1 . Data Collection , 3.2. Data Categorization and 3.3. Data Validation, summarized in Figure 1

3.1. Data Collection

To construct VIVID, we adopted a systematic approach to gather high-quality Vietnamese idioms and proverbs from authoritative lexicographic sources. Our data collection process prioritized linguistic authenticity, cultural representativeness, and semantic diversity.

We compiled idioms and proverbs from two complementary authoritative dictionaries. The primary source was the dictionary by Nguyen Lan (Lân, 2010), which provides a structured and comprehensive collection of widely-used Vietnamese idioms and proverbs accompanied by clear definitions and semantic explanations. This resource is recognized as a standard reference in Vietnamese linguistics and language education, ensuring the reliability and acceptance of the selected expressions

To enhance the linguistic diversity and capture regional variations, we incorporated additional entries from the dictionary compiled by Vu Dung and collaborators (Dung et al., 2000). Unlike Nguyen Lan's work, which predominantly features standardized forms, the Vu Dung dictionary documents numerous idiomatic variants including regional expressions, alternative phrasings, and lexical variations that reflect the richness of Vietnamese idiomatic usage across different dialects and contexts. This dual-source approach enables our dataset to capture both standardized expressions and their naturally occurring variants in Vietnamese communication.

Through this compilation process, we assembled a total of 2,254 idiom-explanation pairs that form the foundation of the VIVID benchmark.

3.2. Data Categorization

3.2.1. Linguistic Complexity Taxonomy

Existing Vietnamese figurative language resources lack systematic categorization with respect to the linguistic and cultural features that challenge computational interpretation. To address this gap, we developed a theory-driven taxonomy grounded in consultation with Vietnamese language experts and informed by preliminary error analysis of LLM outputs.

Our taxonomy organizes Vietnamese idioms and proverbs along three principal aspects and five distinct traits that capture the primary sources of interpretation difficulty, as illustrated in Figure 4

The Meaning Aspect encompasses idioms where semantic interpretation requires understanding beyond literal compositional meaning:

- *Literal expressions*: Proverbs that are purely descriptive without metaphorical layers, yet LLMs frequently over-interpret them by inferring non-existent figurative meanings (e.g., "Da đồng lông móc" describes physical characteristics of buffalo skin but is often misinterpreted as a metaphor for human personality)
- *Pragmatic nuances*: Expressions containing sarcasm, irony, or evaluative connotations that deviate from neutral interpretations (e.g., "Tốt mẽ khoe màu" is intended to mock superficiality but is often misunderstood as promoting humility)

The Lexical Aspect captures vocabulary-related challenges arising from archaic or specialized terminology:

- *Sino-Vietnamese terms*: Expressions containing rare or literary Sino-Vietnamese vocabulary infrequently encountered in contemporary usage (e.g., "Ác khuất non đoài")
- *Uncommon vocabulary*: Idioms featuring archaic or obsolete words no longer common in modern Vietnamese (e.g., "Am thanh cảnh vắng," where "am" refers to a small rural temple)

The Agricultural and Customary Knowledge Aspect includes expressions rooted in traditional Vietnamese cultural practices:

Folk knowledge-based expressions: Proverbs reflecting agricultural practices, seasonal cycles, or customary wisdom from traditional Vietnamese life (e.g., "Cấy tháng Chạp đạp không đổ," a saying based on historical farming experience)

Since idioms may exhibit multiple characteristics simultaneously, entries can be annotated with multiple traits. The distribution of these

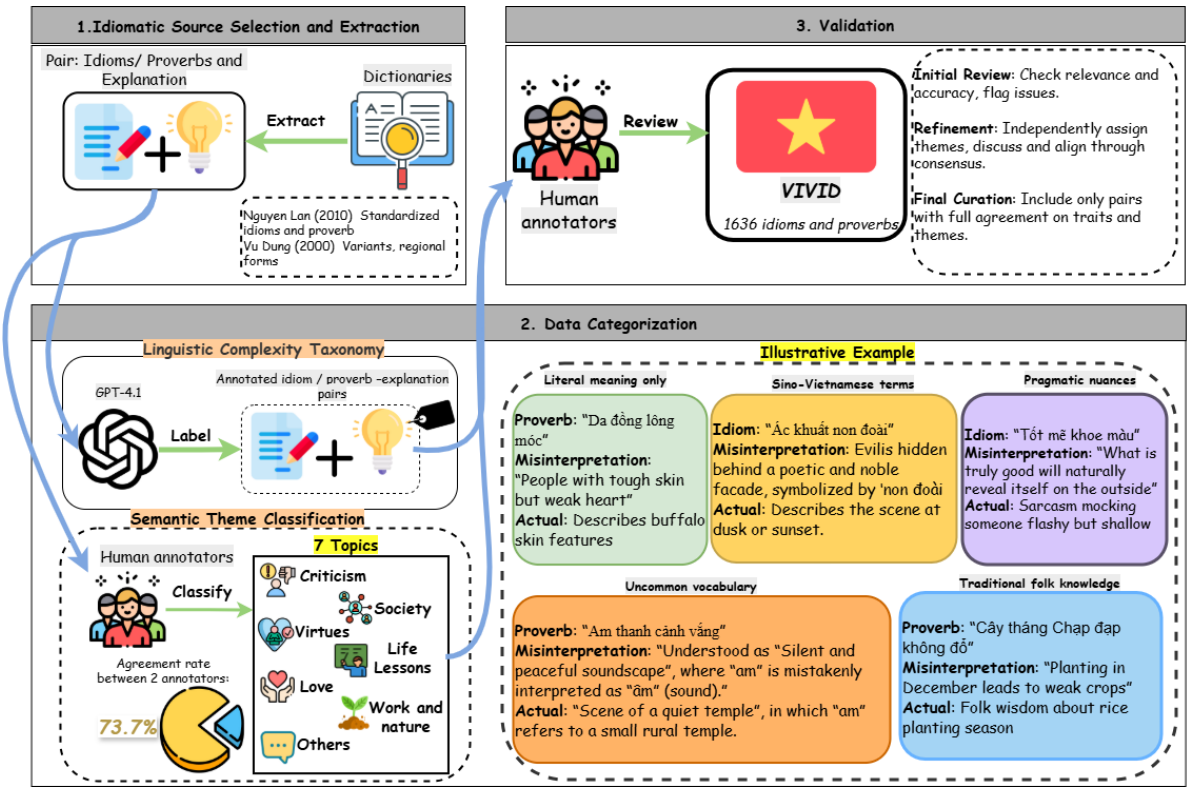

Figure 1: Overview of the data collection and annotation pipeline.

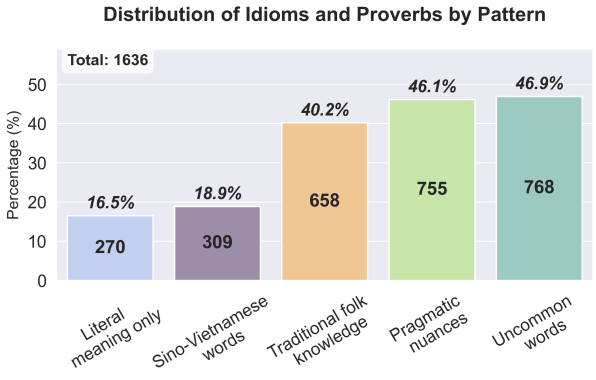

Figure 2: Statistics of VIVID per characteristics.

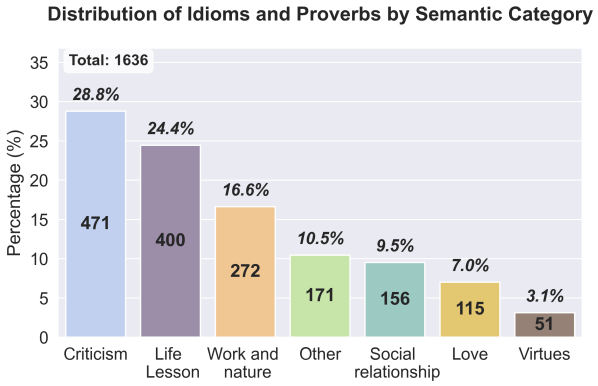

Figure 3: Statistics of VIVID per topics.

complexity characteristics across the dataset is presented in Figure 2, showing that pragmatic nuances (46.9%) and uncommon vocabulary (46.1%) are the most prevalent challenges, followed by Sino-Vietnamese terms (40.2%), folk knowledge-based expressions (18.9%), and literal expressions (16.5%).

3.2.2. Semantic Theme Classification

To facilitate domain-specific evaluation and capture the cultural dimensions of Vietnamese idioms, we categorized each entry according to seven semantic themes that reflect prevalent topics in Vietnamese daily life, folk wisdom, and moral education. This classification schema was derived through thematic analysis of idiomatic content commonly encountered in colloquial usage, folk literature, and educational materials.

The seven themes are defined as follows:

- **Criticism**: Expressions conveying disapproval, mockery, or critical judgment of behaviors, traits, or situations
- **Life Lessons**: Proverbs offering philosophical reflections, moral guidance, or practical wisdom about life
- **Work and Nature**: Idioms related to labor,

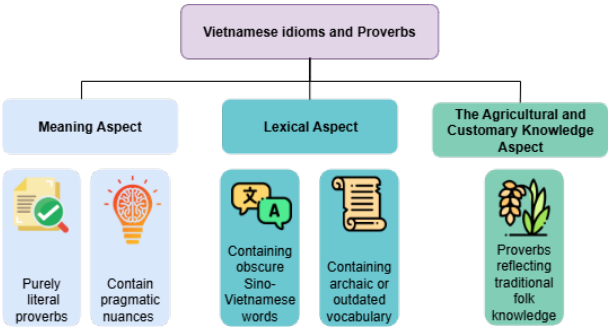

Figure 4: Taxonomy of idiomatic expressions that pose challenges for LLMs due to literal meaning, archaic vocabulary, cultural specificity, and other linguistic features.

agricultural practices, natural phenomena, or traditional livelihoods

- **Other**: Expressions addressing miscellaneous topics including abstract concepts, specific scenarios, or descriptions
- **Social Relationships**: Idioms reflecting interpersonal dynamics within society, including family ties, friendships, neighborhood interactions, and communal relationships
- **Love**: Expressions conveying romantic relationships, emotional bonds, or matters of the heart
- **Virtues**: Idioms emphasizing moral standards, ethical values, or ideal personal qualities

Figure 3 presents the distribution of idioms across these semantic categories, with Criticism (28.8%) being the most represented, followed by Life Lessons (24.4%), Work and Nature (16.6%), Other (10.5%), Social Relationships (9.5%), Love (7.0%), and Virtues (3.1%).

3.3. Data Validation

To ensure the reliability and cultural appropriateness of our annotations, we implemented a rigorous two-stage validation process combining automatic annotation with expert human validation. **Stage 1: Automatic Annotation.** We employed GPT-4.1 to automatically annotate each idiom-explanation pair with one or more labels corresponding to the predefined linguistic complexity traits in our taxonomy (Section 3.2.1). This automated labeling enabled systematic identification of idioms that are particularly prone to misinterpretation due to semantic complexity or cultural specificity, thereby establishing a foundation for subsequent human validation. **Stage 2: Human Validation.** Two native Vietnamese speakers with expertise in Vietnamese linguistics independently reviewed all annotations through a three-step process:

1. *Initial Review*: Each annotator examined the LLM-generated complexity trait labels for contextual relevance and accuracy. Entries with clearly inaccurate or culturally inappropriate labels were flagged for removal, while uncertain cases were marked for collaborative discussion.
2. *Refinement*: For the semantic theme classification (Section 3.2.2), both annotators independently assigned each idiom-explanation pair to one of the seven predefined themes based on its core semantic content. Cases with disagreement or ambiguity were jointly reviewed, and both annotators collaboratively revised the labels to ensure alignment with the predefined taxonomy through consensus-based adjudication.
3. *Final Curation*: All disagreements were resolved through discussion between the two annotators. Only idiom-explanation pairs achieving unanimous agreement on both complexity traits and semantic themes were included in the final dataset.

This validation process resulted in a high-quality curated dataset comprising 1,636 idiom-explanation pairs with validated annotations, providing a reliable foundation for evaluating LLMs' understanding of culturally grounded Vietnamese figurative language.

4. Evaluation Framework and Experiment Setup

In this section, we introduce our figurative language evaluation framework, which integrates both *generative* and discriminative evaluations of Vietnamese idioms and proverbs under the LLM-as-a-Judge paradigm. This setup not only ensures reliable benchmarking but also extends prior evaluation research through prompting strategy analysis, role-based instructions, and human–LLM judge alignment studies.

4.1. Generative Evaluation

To evaluate generative capability, LLMs are asked to explain Vietnamese idioms and proverbs from VIVID. We adopt an open-ended question format rather than constrained multiple-choice, as prior work has shown that in multiple-choice settings, models may rely on superficial patterns rather than genuine understanding (Griot et al., 2025). The quality of each generated explanation is scored

using an LLM-as-a-Judge framework (Gu et al., 2024), with GPT-4.1 serving as an independent evaluator. We test four prompting strategies during evaluation: **(1) Zero-shot**, where the model is given only the explanation and asked to score it; **(2) Demonstration** (Brown et al., 2020), where several annotated examples of high- and low-quality explanations are provided beforehand; **(3) Chain-of-Thought (CoT)** (Wei et al., 2022), in which the model is prompted to reason step-by-step before producing a score; and **(4) Aspect-based Evaluation** (see Figure 7), where the model is instructed to explicitly consider multiple criteria such as factual accuracy, completeness and alignment with common native speaker interpretations.

Each explanation is rated on a scale from 0 to 5 based on its correctness, informativeness, and cultural appropriateness. These scores are then aggregated to compare performance across models and prompting strategies. Prior studies have shown that automatic metrics such as BLEU or COMET fail to capture the nuances of figurative language (Tang et al., 2024a; S. Rezaeimanesh, 2025; Wang et al., 2025b). In contrast, GPT-based evaluations demonstrate high correlation with human judgments, particularly in figurative interpretation tasks (Yang et al., 2025).

Prompt	Mean	Cohen's Kappa
Human Baseline	3.47	1.000
Zero-shot	3.40	0.774
CoT	3.70	0.786
Demonstration	3.44	0.783
Aspect-based	3.49	0.792

Table 1: Evaluation of different LLM-as-a-judge prompting strategies compared to human.

4.1.1. LLM Judge Reliability Analysis

To examine the reliability of the LLM-as-a-Judge approach, we conduct a controlled experiment on 200 Vietnamese idioms and proverbs randomly sampled from our dataset, VIVID. For each entry, GPT-4o generates an explanation under a Zero-shot setting. These outputs are then independently evaluated by GPT-4.1, using each of the four prompting strategies outlined above. Each explanation is rated on three dimensions: semantic accuracy, interpretive completeness, and alignment with native speaker understanding, using a 0–5 scale.

We collect human annotations following the procedure of S. Rezaeimanesh (2025) for comparison with LLM judge. Two native Vietnamese speakers independently annotate 150 idiomatic expressions each, with a 50-sample overlap. Human scores are assigned on the same scale (0–5), achieving strong inter-human annotator reliability.

Type	Model
API-only LLMs	GPT-4o, Gemini 2.5 Flash
Open-source LLMs	**Multilingual:** Llama-4-Scout (109B) Llama-SEA-LION-v3-8B-IT (8B) Qwen-3 (14B) **Vietnamese-specialized:** GreenMind-Medium-14B-R1 (14B) (Tung et al., 2025) Vistral-7B-Chat (7B) (Vo, 2024) VinaLLaMA-7B (7B) (Nguyen et al., 2023)

Table 2: Model selection. The numbers in parentheses indicate the parameters for models.

As reported in Table 1, Aspect-based Evaluation achieves the strongest alignment with human ($\kappa = 0.792$), yielding the closest mean rating (3.49 to 3.47). Consequently, we adapt Aspect-based as the default prompting strategy for automatic scoring of figurative language generative task.

4.2. Discriminative Evaluation

We also conduct multiple-choice classification tasks focusing on (1) topic identification and (2) linguistic characteristic classification, using the prompts illustrated in Figure 9 and Figure 10. These tasks are implemented using the Language Model Evaluation Harness framework (Gao et al., 2024). To ensure a consistent and reproducible evaluation protocol across both opened and closed source models, we adopt exact-match scoring. This choice is motivated by the current limitation of certain closed models (e.g., GPT and Gemini) that do not expose log-likelihoods. All models are evaluated under a 3-shot prompt to ensure understanding of task format and output consistency.

4.3. Benchmark Models

We evaluate a diverse set of open-source and closed-source large language models (LLMs) that vary in scale, language specialization, and accessibility. The open-source models include Llama-4-Scout, Llama-SEA-LION-v3-8B-IT, GreenMind-Medium-14B-R1, Vistral-7B-Chat, VinaLLaMA-7B-Chat, and Qwen-3-14B, while the closed-source models include GPT-4o and Gemini 2.5 Flash (see Table 2 for details).

Two prompting strategies are employed: Zero-shot (Figure 5) and Few-shot prompting (Li et al., 2024). For Few-shot prompting (Figure 6), three examples are provided per task, consisting of two

Category	Model	Prompt	Love	Virtues	Criticism	Work and nature	Society	Life Lessons	Others	Average
Open-source LLMs	Vistral-7B-Chat	Zero-shot	0.25	0.49	0.24	0.34	0.42	0.44	0.39	0.35
		Few-shot	0.54	1.37	0.58	0.69	0.69	0.83	0.70	0.70
	GreenMind-Medium-14B-R1	Zero-shot	1.05	1.40	0.99	0.90	1.04	1.37	1.06	1.09
		Few-shot	0.95	1.12	0.94	0.91	0.86	1.28	0.97	1.02
	Vinallama-7b-chat	Zero-shot	0.13	0.16	0.08	0.11	0.10	0.22	0.10	0.13
		Few-shot	0.23	0.33	0.14	0.16	0.31	0.41	0.17	0.24
	Llama-4-Scout-17B-16E	Zero-shot	0.86	1.27	0.92	0.76	0.88	0.96	1.01	0.91
		Few-shot	1.04	1.61	1.17	1.19	1.24	1.39	1.19	1.24
	Llama-SEA-LION-v3-8B-IT	Zero-shot	0.46	0.80	0.54	0.54	0.50	0.72	0.55	0.58
		Few-shot	0.62	0.61	0.60	0.68	0.56	0.78	0.59	0.65
	Qwen-3-14B	Zero-shot	1.15	1.75	0.95	1.01	1.07	1.28	0.95	1.09
		Few-shot	0.86	1.69	1.01	1.07	1.07	1.37	0.99	1.12
API-only LLMs	Gemini Flash 2.5	Zero-shot	2.51	2.94	2.22	1.76	2.36	2.69	2.03	2.29
		Few-shot	**2.54**	2.86	**2.51**	**2.44**	2.52	**2.89**	**2.46**	2.6
	GPT-4o	Zero-shot	2.33	**3.08**	2.32	2.41	**2.58**	2.60	2.44	2.46
		Few-shot	1.75	2.90	1.95	1.91	1.89	2.31	1.97	2.04

Table 3: Mean score of the models by category (rows: model and prompt style; columns: topic and average). The highest score for each category is in bold. The top-performing open-source models are marked in blue.

Category	Model	Prompt	Only literal expressions	Sino-Vietnamese terms	Uncommon vocabulary	Folk knowledge-based expressions	Pragmatic nuances
Open-source LLMs	Vistral-7B-Chat	Zero-shot	0.40	0.38	0.27	0.34	0.24
		Few-shot	0.76	0.82	0.57	0.74	0.58
	GreenMind-Medium-14B-R1	Zero-shot	0.91	1.30	0.91	1.04	0.97
		Few-shot	0.95	1.20	0.80	1.06	0.94
	Vinallama-7b-chat	Zero-shot	0.07	0.12	0.06	0.15	0.10
		Few-shot	0.12	0.28	0.18	0.26	0.18
	Llama-4-Scout-17B-16E	Zero-shot	0.94	1.18	0.75	0.83	0.83
		Few-shot	1.19	1.50	1.10	1.25	1.18
	Llama-SEA-LION-v3-8B-IT	Zero-shot	0.48	0.74	0.46	0.58	0.54
		Few-shot	0.59	0.79	0.52	0.70	0.58
	Qwen-3-14B	Zero-shot	0.85	1.37	0.92	1.10	1.00
		Few-shot	0.96	1.34	0.92	1.16	1.03
API-only LLMs	Gemini Flash 2.5	Zero-shot	1.70	3.04	2.30	2.15	2.19
		Few-shot	**2.34**	**3.11**	**2.45**	**2.57**	**2.55**
	GPT-4o	Zero-shot	**2.43**	2.86	2.33	2.39	2.44
		Few-shot	1.97	2.36	1.89	2.04	1.92

Table 4: Model performance across five types of linguistic and cultural complexity.

proverbs and one idiom, selected to maximize coverage of the linguistic and cultural characteristics discussed above. This design helps models better infer the task format and expected outputs.

5. Result & Discussion

5.1. Performance on Generative Task

We evaluate eight LLMs on the idiom/proverb explanation generation task in VIVID, under Zero-shot and Few-shot prompting across seven semantic themes (Table 3). Smaller models (7-8B parameters) struggle significantly, with VinaLLaMA-7B-Chat achieving only 0.13 and Vistral-7B-Chat reaching 0.35. Medium-sized models (14B parameters) perform better but remain below 1.5, with both Vietnamese-specialized GreenMind-14B (1.09) and multilingual Qwen-3-14B (1.09) showing comparable performance. This suggests that current language-specific training does not provide clear advantages for cultural figurative understanding at similar model scales. Large-scale models demonstrate substantially higher scores, with GPT-4o (2.46) and Gemini Flash 2.5 (2.29) outperforming all open-source systems, likely due to their extensive pretraining and significantly larger parameter counts.

Nevertheless, all models struggle with idioms/proverbs (best approach with GPT-4o and zero-shot prompting achieves less than 50% correctness on average), underscoring persistent challenges in Vietnamese figurative understanding.

5.1.1. Robustness To Linguistic Complexity

We analyze model robustness across five annotated traits: (1) literal expressions, (2) uncommon vocabulary, (3) rare Sino-Vietnamese terms, (4) folk knowledge–based expressions, and (5) pragmatic nuances.

Table 4 presents the average performance per category, with the Sino-Vietnamese terms category as the least challenging one for LLMs. GPT-4o and Gemini Flash 2.5 outperform all others across these dimensions. GPT-4o achieves 2.86 on Sino-Vietnamese terms and 2.44 on pragmatic nuances (Zero-shot), while Gemini Flash 2.5 (Few-shot) reaches 2.57 on folk knowledge and 2.45 on rare vocabulary. Conversely, lighter models such as VinaLLaMA-7B and Vistral-7B remain below 1.0, highlighting their limited semantic and cultural generalization capacity.

Few-shot prompting generally improves performance across most categories, confirming the benefit of in-context examples for linguistically complex content. The largest gain is seen in Llama-4-Scout-17B ($1.18 \rightarrow 1.50$ on Sino-Vietnamese terms). Notably, GPT-4o shows the opposite trend: its scores drop slightly in Few-shot, suggesting overfitting or semantic rigidity when exposed to structured exemplars. We further analyze this anomaly in Section 5.1.2.

Manual inspection reveals systematic failures: literal over-interpretation (imposing figurative readings where none exist), lexical gaps (mistranslating archaic vocabulary), cultural disconnection (losing contextual meaning in folk-based idioms), and pragmatic flattening (replacing sarcasm or irony with neutral tone). These patterns expose fundamental gaps in current models' cultural competency.

5.1.2. Investigating Prompting Failures in GPT-4o on VIVID

GPT-4o uniquely shows consistent performance decline from zero-shot to few-shot prompting. Analysis reveals GPT-4o overfits to exemplar patterns, imitating moralizing tone rather than focusing on semantic fidelity. For example, "Chưa lại người" (physical recovery after illness) was correctly interpreted in zero-shot but reinterpreted in few-shot as metaphor for immaturity. This indicates didactic examples induce stylistic mimicry and semantic drift, fabricating moral lessons absent in source idioms.

Few-shot prompting also degrades vocabulary handling. "Lang đuôi thì bán, lang trán thì cày" (about buffalo coat patterns) was correctly interpreted under zero-shot but misread under few-shot, where "lang" was mistaken for "sweet potato," producing an unrelated agricultural metaphor. Similar findings by (Phelps et al., 2024) suggest few-shot prompting can hinder figurative interpretation by encouraging stylistic imitation over contextual understanding.

5.2. Evaluating Topic and Linguistic Characteristic Classification Tasks

We evaluate model performance on two discriminative tasks: topic classification and linguistic–cultural taxonomy classification. Accuracy results are summarized in Table 5.

Open-source LLMs show considerable variation across both tasks, while API-based models consistently outperform them, likely due to larger parameter scales and broader pretraining. Among open-weight systems, Llama-4-Scout-17B-16E achieves the highest scores in both topic and linguistic–cultural labeling (0.459 and 0.042), followed by Qwen3-14B (0.447, 0.031) and GreenMind-Medium-14B-R1 (0.455, 0.035). Nevertheless, accuracy on linguistic characteristic classification remains low overall, partly because the exact-match metric penalizes even minor deviations (e.g., added or omitted words). For instance, Viet-Mistral-7B-Chat scored zero in both tasks due to inconsistent adherence to output format and limited capacity (7B parameters). Closed-source models such as GPT-4o and Gemini 2.5 Flash perform best, achieving 0.524 and 0.535 on topic classification, and 0.184 and 0.144 on linguistic–cultural labelling, respectively.

Performance on linguistic–cultural characteristic classification is much lower than on topic classification, reflecting the complexity of culturally grounded expressions. Errors often arise from difficulty interpreting idioms' implicit semantics. For example, "ăn tái ăn tam" ("to eat rare, to eat three times"), literally describing repetitive eating, belongs to the *Other* category but was misclassified by both GPT-4o and Gemini 2.5 Flash as *Criticism*, likely due to reliance on affective cues. Similarly, "bánh giầy nếp cái, con gái họ" ("glutinous rice cake, their daughter"), a proverb metaphorically praising a beautiful girl (*Virtues*), was misclassified by GPT-4o as *Love* and by Gemini 2.5 Flash as Other. These cases illustrate the persistent gap between surface-level lexical cues and the culturally embedded semantics required for accurate classification.

These results demonstrate that state-of-the-art LLMs still struggle to fully capture the meaning and cultural dimensions of Vietnamese idioms and proverbs, underscoring the value of VIVID as a benchmark for advancing culturally aware NLP evaluation.

6. Conclusion

We present VIVID, the first systematic benchmark for evaluating culturally grounded figurative language understanding in Vietnamese, comprising 1,636 idioms and proverbs annotated with linguistic complexity traits and semantic themes. Our extensive evaluation reveals that even state-of-the-art models achieve less than 50% correctness on average, underscoring fundamental gaps in cultural language understanding. Our analysis reveals systematic failure modes including literal over-interpretation, lexical gaps with archaic vocabulary, cultural disconnection in folk-based expressions, and pragmatic flattening. Notably, few-shot prompting does not universally improve performance, with GPT-4o exhibiting stylistic overfitting that degrades semantic fidelity. These findings demonstrate that current LLMs lack the cultural competency required for nuanced figu-

Category	Model	Topic	Linguistics and cultural
Open-source LLMs	Llama-SEA-LION-v3-8B-IT	0.319	0.013
	GreenMind-Medium-14B-R1	0.455	0.035
	Qwen-3-14B	0.447	0.031
	Llama-4-Scout-17B-16E	0.459	0.042
	Vinallama-7B-Chat	0.079	0.016
	Vistral-7B-Chat	0.000	0.000
API-based LLMs	GPT-4o	0.524	**0.184**
	Gemini Flash 2.5	**0.535**	0.144

Table 5: Model accuracy across topic and linguistic-cultural classification tasks.

rative interpretation despite their general linguistic capabilities. VIVID establishes a robust evaluation framework for Vietnamese figurative language comprehension and provides actionable insights for developing culturally aware NLP systems that go beyond surface-level pattern matching to achieve genuine figurative language understanding.

7. Limitations

First, our automatic evaluation framework relies on a proprietary closed-source model (GPT-4.1). Dependence on a closed model raises concerns regarding transparency and long-term reproducibility. Commercial models may be updated or modified over time without full disclosure of technical details, potentially leading to changes in evaluation behavior in future versions. This may affect the stability of results if experiments are replicated at different points in time. Second, the reliability analysis between the LLM judge and human annotators is conducted on a subset of 200 samples due to the substantial cost and effort associated with manual annotation. While this sample size is adequate for estimating agreement trends between evaluators, it may not fully capture the complete diversity of linguistic forms, rarity levels, and cultural nuances present in VIVID. Future work may extend validation to a larger scale and compare multiple evaluation models, including open-source alternatives, to further enhance the transparency and reproducibility of the proposed evaluation framework.

8. Acknowledgements

This publication has emanated from research supported in part by grants from Research Ireland under Grant [12-RC-2289-P2] and [18/CRT/6223] which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

9. Bibliographical References

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266, Albuquerque, New Mexico. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Brown, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluation? *arXiv preprint arXiv:2305.01937*.

Phong Nguyen-Thuan Do, Son Quoc Tran, Phu Gia Hoang, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. VLUE: A new benchmark and multi-task knowledge transfer learning for Vietnamese natural language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 211–222, Mexico City, Mexico. Association for Computational Linguistics.

Vũ Dung et al. 2000. *Từ điển thành ngữ & tục ngữ Việt Nam [Dictionary of Vietnamese Idioms & Proverbs]*. Nhà xuất bản Văn Hóa Thông Tin, Hà Nội.

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, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2025. Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5321–5341, Vienna, Austria. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Vong Anh Ho, Duong Nguyen, Danh Hoang Thanh Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2019a. Emotion recognition for vietnamese social media text. *ArXiv*, abs/1911.09339.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2019b. Emotion recognition for Vietnamese social media text. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING 2019)*.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022a. Vinli: A vietnamese corpus for studies on open-domain natural language inference. In *International Conference on Computational Linguistics*.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022b. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean.

Khang Le, Hien Nguyen, Tung Le Thanh, and Minh Nguyen. 2022a. VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6521–6529, Marseille, France. European Language Resources Association.

Khang Le, Hien Nguyen, Tung Le Thanh, and Minh Nguyen. 2022b. VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6521–6529, Marseille, France. European Language Resources Association.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022a. 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*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. 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*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Nguyễn Lân. 2010. *Từ điển thành ngữ, tục ngữ Việt Nam [Dictionary of Vietnamese Idioms and Proverbs]*. Nhà xuất bản Văn Hóa Thông Tin, Hà Nội.

Mubarak Banu Sayed Naziya Shaikh, Jyoti Pawar. 2024. Konidioms corpus: A dataset of idioms in konkani language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Quan Nguyen, Huy Pham, and Dung Dao. 2023. Vinallama: Llama-based vietnamese foundation model. Technical Report arXiv:2312.11011, arXiv.

Sang Quang Nguyen and Kiet Van Nguyen. 2025. A large-scale benchmark for Vietnamese sentence paraphrases. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1045–1060, Albuquerque, New Mexico. Association for Computational Linguistics.

Ivana Filipović Petrović, Miguel López Otal, and Slobodan Beliga. 2024. Croatian idioms integration: Enhancing the lidioms multilingual linked idioms dataset.

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

Hoàng Phê. 2021. *Từ điển Tiếng Việt [Vietnamese Dictionary]*. Nhà xuất bản Hồng Đức, Hà Nội.

Yadollah Yaghoobzadeh S. Rezaeimanesh, Faezeh Hosseini. 2025. Large language models for persian ←→ english idiom translation. In *Proceedings of the 2025 North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024a. Creative and context-aware translation of east Asian idioms with GPT-4. In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 1–21, Miami, Florida, USA. Association for Computational Linguistics.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024b. Creative and context-aware translation of East Asian idioms with GPT-4. In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 1–21, Miami, Florida, USA. Association for Computational Linguistics.

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

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dien Dinh. 2024. ViGLUE: A Vietnamese general language understanding benchmark and analysis of Vietnamese language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4174–4189, Mexico City, Mexico. Association for Computational Linguistics.

Luu Quy Tung, Hoang Quoc Viet, Pham Bao Loc, and Vo Trong Thu. 2025. Greenmind: A next-generation vietnamese large language model for structured and logical reasoning. Technical Report arXiv:2504.16832, arXiv.

James Vo. 2024. Vi-mistral-x: Building a vietnamese language model with advanced continual pre-training. Technical Report arXiv:2403.15470, arXiv.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. Technical Report arXiv:1905.00537, arXiv.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. Technical Report arXiv:1804.07461, arXiv.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025a. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662,

Vienna, Austria. Association for Computational Linguistics.

Minghan Wang, Viet-Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025b. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Pacific Asia Conference on Language, Information and Computation.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhen Xu, Xinjin Li, Yingqi Huan, Veronica Minaya, and Renzhe Yu. 2025. From course to skill: Evaluating llm performance in curricular analytics. *arXiv preprint arXiv:2505.02324*.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025. Evaluating llms on chinese idiom translation. In *Conference on Language Modeling (COLM)*.

Yuhao Zeng, Zongxi Li, Ruoxi Cui, Linmei Hu, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Revisiting large language models hallucination in instruction tuning: An empirical study. Technical Report arXiv:2504.05747, arXiv.

Hengran Zhang, Hu Xu, and Mu Li. 2024. Are large language models good at utility judgments? *arXiv preprint arXiv:2403.19216*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

A. Prompts

A.1. Generative Task Prompts

We present the exact prompt templates used for explanation generation. All prompts are issued in Vietnamese to ensure the models respond in the target language and leverage their Vietnamese-language capabilities.

Zero-shot Prompt. The zero-shot prompt (Figure 5) is intentionally minimal, providing only the idiom or proverb and asking for a concise Vietnamese explanation.

Vietnamese

Giải thích ngắn gọn bằng tiếng Việt ý nghĩa câu thành ngữ sau: `{phrase}`

English (translation)

Briefly explain in Vietnamese the meaning of the following idiom/proverb: `{phrase}`

Figure 5: Zero-shot prompt for idiom explanation generation (Vietnamese original left, English translation right).

Few-shot Prompt. The few-shot prompt (Figure 6) situates the model as a Vietnamese language expert and provides three demonstrative examples covering diverse idiom types (a social proverb, a culinary proverb, and an ironic proverb) before asking for the target explanation.

A.2. LLM-as-a-Judge Prompt

The aspect-based judge is shown in Figure 7. It instructs GPT-4.1 to evaluate a generated explanation across four explicit dimensions before producing a single overall similarity score on a 0–5 scale. Crucially, a score of 0 on Criterion 1 (core-meaning accuracy) forces an overall score of 0 regardless of other criteria, preventing superficially fluent but semantically incorrect responses from receiving inflated scores.

A.2.1. Judge Prompt In-Context Examples

Two annotated examples are embedded in the judge prompt (see Figure 8) to calibrate the scoring scale: one high-quality response (score 5) and one incorrect response (score 1).

B. Discriminative Task Prompts

The discriminative evaluation is implemented via the Language Model Evaluation Harness framework (Gao et al., 2024) using exact-match scoring under 3-shot prompting. All prompts are in Vietnamese.

Vietnamese

Bạn là chuyên gia ngôn ngữ tiếng Việt. Nhiệm vụ của bạn là **giải thích bằng tiếng Việt** ý nghĩa của câu thành ngữ hoặc tục ngữ sau một cách ngắn gọn, rõ ràng và dễ hiểu. Hãy đọc và tham khảo các cách giải thích dưới đây:

Ví dụ 1: "dân chỉ công bộc"
Giải thích: Câu tục ngữ "dân chỉ công bộc" có ý nghĩa là dân chúng chính là người chủ, còn những người làm việc trong bộ máy nhà nước, chính quyền là "công bộc", tức là người phục vụ cho dân. Thành ngữ này nhấn mạnh vai trò quan trọng của người dân trong xã hội và trách nhiệm của những người làm công tác quản lý, lãnh đạo trong việc phục vụ và đáp ứng nhu cầu của nhân dân.

Ví dụ 2: "Cần tái cải nhừ"
Giải thích: Một kinh nghiệm ăn uống: với rau cần khi thả vào nước sôi, đun cho sôi lại là bắc ra ngay (tái) để rau không bị nhũn và tránh có mùi nồng; với rau cải phải đun kỹ mới hết hăng hết cay.

Ví dụ 3: "ăn cho thủng nổi trôi rế"
Giải thích: diễn tả việc ăn uống một cách rất nhiều, không kiêng khem hay tiết chế. Ý nghĩa của câu này là chỉ sự ăn uống quá mức, đến mức có thể "thủng" cả cái rế (vật dụng đựng thức ăn). Chê những kẻ ăn uống tham lam quá độ.

Bây giờ, hãy giải thích câu sau: `{phrase}`
giải thích:

English (translation)

You are a Vietnamese language expert. Your task is to **explain in Vietnamese** the meaning of the following idiom or proverb concisely, clearly, and accessibly. Please read and refer to the example explanations below:

Example 1: "dân chỉ công bộc"
Explanation: The proverb "dân chỉ công bộc" means that the people are the true masters, while those working in government and state apparatus are "công bộc" — servants of the people. It emphasises the vital role of citizens in society and the responsibility of officials to serve and meet the needs of the populace.

Example 2: "Cần tái cải nhừ"
Explanation: A culinary tip: water celery (cần) should be removed from boiling water as soon as it returns to a boil (tái) to stay crisp and avoid a pungent smell; mustard greens (cải) must be cooked thoroughly to lose their bitterness and sharpness.

Example 3: "ăn cho thủng nổi trôi rế"
Explanation: Depicts eating in extreme excess without any restraint. The saying criticises gluttonous behaviour so severe it could "pierce through" the rế (a vessel used to hold food). It mocks those who are insatiably greedy eaters.

Now, please explain the following idiom/proverb: `{phrase}`
explanation:

Figure 6: Few-shot prompt for idiom explanation generation, with three in-context demonstrations (Vietnamese original left, English translation right).

B.1. Topic Classification Prompt

The topic classification prompt is shown in Figure 9. Models are asked to identify which of seven semantic themes best describes a given idiom or proverb, returning a single integer (1–7).

B.2. Linguistic Complexity Classification Prompt

The linguistic complexity classification prompt is presented in Figure 10. Models are asked to identify *all applicable* complexity traits for a given idiom or proverb, returning a comma-separated list of integers in ascending order (e.g., `1, 3, 5`). An empty string is returned when no trait applies. Exact-match scoring penalises any deviation from the gold label, including reordering or additional characters.

C. Model and Hyperparameter Details

For generation, all models use `temperature=0.7` (default) and `max_tokens=150`. For judging, GPT-4.1 uses `temperature=0` and `max_tokens=10`.

Vietnamese

[Hướng dẫn]
Bạn là chuyên gia ngôn ngữ. Nhiệm vụ của bạn là đánh giá độ tương đồng giữa giải thích gốc (con người) và giải thích của LLM cho một câu thành ngữ/tục ngữ, dựa trên các tiêu chí sau:

1. **Độ chính xác của ý nghĩa cốt lõi (0–5 điểm):** Giải thích của LLM nắm bắt đúng và chính xác ý nghĩa chính của câu thành ngữ/tục ngữ so với giải thích gốc? (5: Rất chính xác, 0: Sai lệch hoàn toàn)
2. **Khả năng nắm bắt sắc thái, ngữ cảnh, ẩn ý (0–5 điểm):** Giải thích của LLM có phản ánh được các sắc thái, ngữ cảnh sử dụng điển hình, hoặc ý nghĩa ẩn sâu của câu nói như giải thích gốc không? (5: Rất tương đồng, 0: Không phản ánh được)
3. **Chất lượng diễn đạt, sự rõ ràng, lưu loát (0–5 điểm):** Giải thích của LLM có rõ ràng, dễ hiểu, mạch lạc và tự nhiên không? (5: Rất tốt, 0: Rất kém/khó hiểu)
4. **Độ đầy đủ (0–5 điểm):** Giải thích của LLM có bao gồm các ý chính được đề cập trong giải thích gốc không? (5: Rất đầy đủ, 0: Thiếu nhiều ý chính)
5. **Độ tương đồng tổng thể (0–5 điểm):** Tổng hợp các đánh giá trên, mức độ tương đồng chung là bao nhiêu? (5: Rất tương đồng, 0: Không tương đồng)

[In-context examples — see Figure 8]

[Nhiệm vụ cần đánh giá]
Thành ngữ/Tục ngữ: `{phrase}`
Giải thích của Con người: `{ground_truth}`
Giải thích của LLM: `{llm_output}`

Đánh giá chi tiết:
1. Độ chính xác của ý nghĩa cốt lõi:
2. Khả năng nắm bắt sắc thái, ngữ cảnh, ẩn ý:
3. Chất lượng diễn đạt, sự rõ ràng, lưu loát:
4. Độ đầy đủ:
Nếu tiêu chí 1 được 0 điểm thì Độ tương đồng tổng thể được 0.
Độ tương đồng tổng thể: hãy suy nghĩ sau đó chỉ trả về duy nhất đánh giá của độ tương đồng tổng thể và không giải thích gì thêm

English (translation)

[Instructions]
You are a language expert. Your task is to evaluate the similarity between the original (human-written) explanation and the LLM-generated explanation for a given Vietnamese idiom or proverb, based on the following criteria:

1. **Accuracy of core meaning (0–5 points):** Does the LLM explanation correctly and precisely capture the main meaning of the idiom/proverb compared to the original explanation? (5: Very accurate, 0: Completely incorrect)
2. **Capture of nuance, context, and implicit meaning (0–5 points):** Does the LLM explanation reflect the nuances, typical usage contexts, or deeper implied meanings of the expression as in the original explanation? (5: Highly similar, 0: Does not reflect them at all)
3. **Expression quality, clarity, and fluency (0–5 points):** Is the LLM explanation clear, easy to understand, coherent, and natural? (5: Excellent, 0: Very poor/difficult to understand)
4. **Completeness (0–5 points):** Does the LLM explanation include the key points mentioned in the original explanation? (5: Very complete, 0: Missing many key points)
5. **Overall similarity (0–5 points):** Synthesising the above assessments, what is the overall degree of similarity? (5: Highly similar, 0: Not similar at all)

[In-context examples — see Figure 8]

[Task to evaluate]
Idiom/Proverb: `{phrase}`
Human explanation: `{ground_truth}`
LLM explanation: `{llm_output}`

Detailed evaluation:
1. Accuracy of core meaning:
2. Capture of nuance, context, and implicit meaning:
3. Expression quality, clarity, and fluency:
4. Completeness:
If criterion 1 receives 0 points, the overall similarity score must also be 0.
Overall similarity: Think carefully, then return only the overall similarity score and provide no further explanation.

Figure 7: Aspect-based LLM-as-a-Judge prompt (Vietnamese original left, English translation right). GPT-4.1 outputs only the integer overall similarity score (0–5).

Vietnamese

Ví dụ Minh họa 1
Thành ngữ/Tục ngữ: *Chưa đủ lông đủ cánh*
Giải thích nghĩa gốc: Còn non nớt, chưa đủ khả năng tự lập.
Giải nghĩa do LLM: Câu thành ngữ "Chưa đủ lông đủ cánh" được dùng để chỉ một người hoặc một sự vật, sự việc nào đó chưa phát triển hoàn thiện, chưa đủ trưởng thành hoặc chưa có đủ khả năng, kinh nghiệm để thực hiện một nhiệm vụ, công việc nào đó một cách độc lập.
Đánh giá chi tiết:
1. Độ chính xác của ý nghĩa cốt lõi: 5/5
2. Khả năng nắm bắt sắc thái, ngữ cảnh, ẩn ý: 5/5
3. Chất lượng diễn đạt, sự rõ ràng, lưu loát: 5/5
4. Độ đầy đủ: 5/5
Độ tương đồng tổng thể: 5/5

Ví dụ Minh họa 2
Thành ngữ/Tục ngữ: *Kim chỉ có đầu*
Giải thích nghĩa gốc: Phải có trên dưới, đầu đuôi, trật tự văn phép.
Giải nghĩa do LLM: Câu thành ngữ "Kim chỉ có đầu" mang ý nghĩa rằng mọi việc đều có nguyên nhân và nguồn gốc, không có gì tự nhiên mà xảy ra.
Đánh giá chi tiết:
1. Độ chính xác của ý nghĩa cốt lõi: 1/5
2. Khả năng nắm bắt sắc thái, ngữ cảnh, ẩn ý: 1/5
3. Chất lượng diễn đạt, sự rõ ràng, lưu loát: 3/5
4. Độ đầy đủ: 1/5
Độ tương đồng tổng thể: 1/5

English (translation)

Illustrative Example 1
Idiom/Proverb: *Chưa đủ lông đủ cánh* ("Not yet fully feathered")
Human explanation: Still inexperienced; not yet capable of independence.
LLM explanation: The idiom "Chưa đủ lông đủ cánh" is used to describe a person or thing that has not yet fully developed, is not sufficiently mature, or lacks the ability and experience to perform a task or role independently.
Detailed scores:
1. Accuracy of core meaning: 5/5
2. Capture of nuance, context, and implicit meaning: 5/5
3. Expression quality, clarity, and fluency: 5/5
4. Completeness: 5/5
Overall similarity: 5/5

Illustrative Example 2 (low-quality response)
Idiom/Proverb: *Kim chỉ có đầu* ("The needle and thread have a head")
Human explanation: There must be a proper order — top and bottom, beginning and end, social hierarchy and etiquette.
LLM explanation: The idiom "Kim chỉ có đầu" means that everything has a cause and origin; nothing happens spontaneously.
Detailed scores:
1. Accuracy of core meaning: 1/5
2. Capture of nuance, context, and implicit meaning: 1/5
3. Expression quality, clarity, and fluency: 3/5
4. Completeness: 1/5
Overall similarity: 1/5

Figure 8: In-context scoring examples embedded in the judge prompt to calibrate the 0–5 scale (Vietnamese original left, English translation right).

Vietnamese

Bạn là chuyên gia ngôn ngữ. Dưới đây là một câu thành ngữ/tục ngữ: "`{phrase}`"
Nhiệm vụ của bạn là xác định câu này thuộc một trong các đặc điểm sau:

1. **Tình yêu đôi lứa và hôn nhân**
Ví dụ: *chẳng nên tình trước nghĩa sau, có con ta gả cho nhau thiệt gì*
2. **Ca ngợi phẩm chất và đạo đức con người**
Ví dụ: *cân quắc anh hùng*
3. **Phê phán và châm biếm các thói quen xấu, hành vi tiêu cực**
Ví dụ: *ai biết uốn câu cho vừa miệng cá*
4. **Kinh nghiệm sống về sản xuất, lao động và thiên nhiên**
Ví dụ: *ác tắm thì ráo, sáo tắm thì mưa*
5. **Các mối quan hệ xã hội, gia đình**
Ví dụ: *ắm con chồng hơn bồng cháu ngoại*
6. **Triết lý và bài học về cách sống**
Ví dụ: *con ruồi đậu nặng đồng cân*
7. **Các chủ đề khác** (nếu không phù hợp với các mục trên)

Chỉ trả lời duy nhất một số (1–7). Không giải thích, không thêm ký tự nào khác.

English (translation)

You are a language expert. Below is a Vietnamese idiom or proverb: "`{phrase}`"
Your task is to identify which of the following categories best describes it:

1. **Romantic love and marriage**
Example: *chẳng nên tình trước nghĩa sau, có con ta gả cho nhau thiệt gì*
2. **Praising human virtues and moral character**
Example: *cân quắc anh hùng*
3. **Criticising and satirising bad habits or negative behaviour**
Example: *ai biết uốn câu cho vừa miệng cá*
4. **Life experience about production, labour, and nature**
Example: *ác tắm thì ráo, sáo tắm thì mưa*
5. **Social and family relationships**
Example: *ắm con chồng hơn bồng cháu ngoại*
6. **Philosophy and life lessons**
Example: *con ruồi đậu nặng đồng cân*
7. **Other topics** (if none of the above apply)

Reply with a single number only (1–7). No explanation, no extra characters.

Figure 9: Topic classification prompt for the `topic_labeling` discriminative task (Vietnamese original left, English translation right).

Vietnamese

Bạn là chuyên gia ngôn ngữ. Dưới đây là một câu thành ngữ/tục ngữ: "`{phrase}`"
Nhiệm vụ của bạn là xác định câu này thuộc **những đặc điểm nào** trong danh sách dưới đây. Kết quả trả về là danh sách số thứ tự đặc điểm, cách nhau bằng dấu phẩy (ví dụ: "1,3,5"). Nếu không thuộc đặc điểm nào, trả về chuỗi rỗng "".

Danh sách đặc điểm:
1. Là câu tục ngữ **không mang nghĩa bóng** (chỉ mang nghĩa đen/kỹ thuật).
Ví dụ: *Da đồng lông móc* (kinh nghiệm chọn trâu, không hàm ý sâu xa).
2. Có **từ Hán Việt** xuất hiện.
Ví dụ: *Nhất tự vi sư, bán tự vi sư*.
3. Có từ **ít phổ biến, ít dùng** trong ngôn ngữ hiện đại.
Ví dụ: *Am thanh cảnh vắng* (từ "am" nghĩa là chùa nhỏ).
4. Thể hiện **kinh nghiệm dân gian xưa** của người Việt Nam.
Ví dụ: *Cấy thưa thừa thóc, cấy mau dốc bồ*.
5. Mang sắc thái **mỉa mai, châm biếm hoặc tiêu cực**.
Ví dụ: *Tốt mẽ khoe màu*.

Chỉ trả lời bằng các số đặc điểm, sắp xếp theo thứ tự tăng dần, cách nhau bằng dấu phẩy. Không giải thích, không thêm ký tự nào khác ngoài danh sách số.

English (translation)

You are a language expert. Below is a Vietnamese idiom or proverb: "`{phrase}`"
Your task is to identify **all applicable characteristics** from the list below. Return a comma-separated list of characteristic numbers in ascending order (e.g., "1,3,5"). If none apply, return an empty string "".

List of characteristics:
1. The proverb **carries no figurative meaning** (literal/technical meaning only).
Example: *Da đồng lông móc* (buffalo selection criteria, no deeper implication).
2. Contains **Sino-Vietnamese vocabulary**.
Example: *Nhất tự vi sư, bán tự vi sư*.
3. Contains words **rarely used** in modern Vietnamese.
Example: *Am thanh cảnh vắng* ("am" means a small rural temple).
4. Reflects **traditional Vietnamese folk knowledge**.
Example: *Cấy thưa thừa thóc, cấy mau dốc bồ*.
5. Carries a **sarcastic, ironic, or negative tone**.
Example: *Tốt mẽ khoe màu*.

Reply only with the characteristic numbers in ascending order, comma-separated. No explanation, no extra characters beyond the number list.

Figure 10: Linguistic complexity classification prompt for the `pattern_labeling` discriminative task (Vietnamese original left, English translation right). Models must predict all applicable traits; exact match is required for a correct prediction.