FLUID QA: A Multilingual Benchmark for Figurative Language Usage in Dialogue across English, Chinese, and Korean

Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An,
Xiaonan Wang, Gyuri Choi, Hansaem Kim[†]
Yonsei University, South Korea
{seoyoon.park, choihz519, minseonk1m, aaasb9946,
nan, gyuri1345, khss}@yonsei.ac.kr

Abstract

Figurative language conveys stance, emotion, and social nuance, making its appropriate use essential in dialogue. While large language models (LLMs) often succeed in recognizing figurative expressions at the sentence level, their ability to use them coherently in conversation remains uncertain. We introduce FLUID QA, the first multilingual benchmark that evaluates figurative usage in dialogue across English, Korean, and Chinese. Each item embeds figurative choices into multi-turn contexts. To support interpretation, we include FLUTE-bi, a sentence-level diagnostic task. Results reveal a persistent gap: models that perform well on FLUTE-bi frequently fail on FLUID QA, especially in sarcasm and metaphor. These errors reflect systematic rhetorical confusion and limited discourse reasoning. FLUID QA provides a scalable framework for assessing usage-level figurative competence across languages.

1 Introduction

Figurative language, defined as the use of at least one lexical item in a nonliteral or nonstandard sense (Paul, 1970), is a core component of everyday communication (Gibbs, 1994). Producing and interpreting figurative expressions requires not only semantic knowledge but also sensitivity to context and pragmatic appropriateness (Fodor & Katz, 1964; Roberts & Kreuz, 1994). As such, evaluating large language models (LLMs) for figurative competence demands more than recognition of figurative markers in isolation, it requires assessment in situated, communicative settings. Despite this, most existing benchmarks focus on sentence-level classification or inference tasks (Zheng et al., 2019; Chakrabarty et al., 2021; Liu et al., 2024), which assess recognition ability without capturing discourse-level usage or contextual fit. This limitation is particularly concerning given the cross-linguistic and culturally embedded nature of figurative language.

To address these gaps, we introduce FLUID QA, the first multilingual benchmark designed to evaluate figurative language usage within dialogue. FLUID QA situates figurative expression selection within multi-turn conversational contexts, testing whether LLMs can make pragmatically appropriate choices across English, Korean, and Chinese. To aid interpretation of usage-level failures, we additionally present FLUTE-bi, a lightweight diagnostic set targeting sentence-level recognition. Rather than constructing data from scratch, we repurpose and restructure the existing FLUTE dataset (Chakrabarty et al., 2022), leveraging its high-quality figurative instances across rhetorical categories. This approach enables us to evaluate not only whether LLMs can recognize figurative expressions, but also whether they can deploy them coherently in real-world dialogue across languages. Our contributions are threefold:

(1) We propose FLUID QA, the first benchmark to assess figurative language usage in dialogue across multiple languages and rhetorical categories.
(2) We provide a cross-linguistic analysis of recognition and usage divergence, showing how pragmatic failure varies by category and mod els.

[†] Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30280–30294
November 4-9, 2025 ©2025 Association for Computational Linguistics

Per language	**Simile**	**Metaphor**	**Idiom**	**Sarcasm**	**Sum**
FLUTE-bi	200	200	200	200	800
FLUID QA	200	200	200	200	800

Table 1: FLUTE-bi and FLUID QA dataset statistics. FLUTE-bi dataset consists of literal-figurative sentence pairs and FLUID QA has single dialogue QA per instance.

Type	**Cultural Adaptation Examples**
Word-to-word correction	EN) The republicans are floating the idea of a tax reform. - metaphor KO) 공화당원들이 세제 개혁 아이디어를 띄우고 있다. (The republicans are flying the idea of a tax reform..) → ‘float the idea’ could be metaphor in English, but the Korean translation lost metaphor meaning cause ‘float’ and ‘idea’ are not often collocate each other in Korean. Correct-KO) 공화당원들이 세제 개혁 아이디어를 **짜내는 중**이다 (The republicans are ***squeezing*** **the idea** of a tax reform.)
Cultural habits	EN) It's really awesome how my family didn't bother to show up for my kids 6th birthday party. - sarcasm ZH) 我家人没来参加我孩子的六岁生日派对，很好。 (My family didn't come to my kid's sixth birthday party, pretty good.) → Need to add an appropriate sarcastic tone to match Chinese emotion expression habits Correct-ZH) 我家人没来参加我孩子的六岁生日派对，**真是太棒了**。 (My family didn't come to my kid's sixth birthday party which was **really awesome**!)

Table 2: Examples of Cultural Adaptation Translation. One example per language is provided for illustration, but all types of error in both languages were corrected by cultural adaptation prompt.

(3) We uncover systematic error patterns including category-specific confusion and stance misinterpretation that reveal structural limitations in current LLMs’ discourse reasoning.

By reframing figurative competence as a context-sensitive, usage-level ability, FLUID QA exposes a persistent blind spot in current LLMs’ communicative reasoning. Together with FLUTE-bi, it offers a layered framework for diagnosing figurative understanding in multilingual dialogue settings.

2 Related Works

LLMs still struggle with multilingual support due to the dominance of English in resources (Ahuja et al.,2023; Ahuja et al.,2024; Nicholas & Bhatia, 2023; Dong et al., 2024). Figurative language studies focus primarily on English, emphasizing sentence-level classification or inference tasks (Chakrabarty et al., 2021, 2022; Liu et al., 2022; Stowe et al., 2022; Jang et al., 2023). Multilingual studies follow similar structures (Lai et al., 2022; Kabra et al., 2023), and while some attempt QA or cloze-style tasks (Zheng et al., 2019; Rakshit et al., 2022), they remain sentence-based, limiting conversational applicability. While some recent studies evaluate figurative understanding in dialogue (Jhamtani et al., 2021; Settaluri et al., 2024), they remain English-only, whereas our work extends this line to pragmatically grounded, multilingual dialogue evaluation.

With LLM advancements, prompt-based translation has become common including for figurative language (Yamada, 2023; Son et al., 2024; Rezaeimanesh et al., 2024; Khoshtab et al., 2024; Donthi et al., 2025). Studies confirm its effectiveness in improving translation quality and cultural nuance adaptation (Gao et al., 2024; Tang et al., 2024; Singh et al., 2024; He et al., 2024).

This work advances prior research by proposing a multilingual, dialogue-level benchmark with culturally adapted translations, addressing the English-only and sentence-level focus of existing studies.

3 Dataset Construction

Figurative language is sparse and culturally grounded, making it difficult to evaluate in low-resource or cross-lingual contexts. Instead of creating new data from scratch, we build on the FLUTE dataset, which offers high-quality English examples across rhetorical types. We sample 200 instances per category (idiom, metaphor, simile, sarcasm) to construct parallel data. This approach. allows us to focus on contextual and multilingual alignment without constructing from the ground up.

EN	Replaced Idiom
Make money hand over fist.	KO 돈방석에 앉다. (Sitting on the money seat.)
In cold blood.	ZH 袖手旁观(to stand by and watch without taking any action.)

Table 3: Examples of culturally equivalent idioms used in translation. Replacements were selected to preserve semantic and pragmatic alignment across languages.

Section 3.1 outlines our culturally adaptive translation pipeline. Section 3.2 introduces FLUID QA, a usage-level benchmark in multi-turn dialogue. Section 3.3 presents FLUTE-bi, a sentence-level task targeting recognition. Dataset statistics are summarized in Table 1.

3.1 Cultural Adaptation for Multilingual Construction

Figurative language is shaped by cultural norms, making direct translation unreliable for cross-lingual evaluation. Literal translations often miss figurative meaning or conflict with cultural language use. To ensure cross-linguistic validity, we adopted a culturally adaptive prompting strategy using GPT-4o for Korean and Chinese translations (Table 2). Following Lai et al. (2023), who found that prompt language has minimal effect on output quality, we used English prompts based on He et al. (2024)'s 'Translator' persona and Singh et al. (2024)'s Cultural Adaptation Prompt (Appendix A). For idioms, which are syntactically fixed and culturally specific (Sprenger, 2003; Knappe, 2012), we replaced them with culturally equivalent idioms in Korean and Chinese to preserve both semantic and pragmatic meaning. Examples of replacements are in Table 3.

We conducted pairwise preference comparisons between literal translations and culturally adapted versions using the Bradley-Terry model (Bradley & Terry, 1952). For each target language (Korean and Chinese), native speakers participated in the evaluation. The results consistently favored the culturally adapted translations, showing statistically significant improvements over literal counterparts ($p < 0.001$). In addition to subjective preference, we also examined the downstream task performance under each translation condition. Models consistently performed better when trained and evaluated on culturally adapted versions, suggesting that literal translations may introduce subtle mismatches or noise. (See Appendix C for details.)

As a final step to ensure the highest quality, all translations were manually reviewed and post-edited by native-speaking authors for syntactic fluency and cultural compatibility.

3.2 FLUID QA: Contextual Figurative Usage Benchmark

FLUID QA is our primary benchmark, designed to assess discourse-level figurative competence. Each item presents a short multi-turn dialogue (3–4 turns) ending in a cloze-style prompt, where the model selects the most contextually appropriate figurative expression from four candidates: a pragmatically correct answer, a semantically similar distractor, an unrelated option, and an incongruent or antonymic distractor.

This setup probes pragmatic reasoning, including sensitivity to tone, speaker intent, and discourse-level appropriateness. We conceptualize this ability as 'figurative usage', distinct from recognition tasks that simply label isolated sentences. Usage entails selecting expressions that align with contextual nuance and social meaning across dialogue turns, reflecting applied communicative reasoning rather than surface-level recognition.

To operationalize this, we adopt a multiple-choice format. This offers a balance between the simplicity of classification task and the uncontrolled variability of free-form generation, enabling both expressive challenge and evaluation stability. The format also mirrors real-world language proficiency tests (e.g., SAT, TOEFL), where pragmatic competence is commonly assessed through structured choices.

Data construction was guided by FLUTE's literal, figurative, and explanatory annotations. We generated items using GPT-4o with teacher-style prompting inspired by educational cloze tests (Xie et al., 2018). Full prompt details are provided in Appendix B, and all outputs were post-edited for fluency and coherence. The final dataset comprises 800 QA items per language (English, Chinese, Korean), evenly distributed across four rhetorical

Parallel_id	Lang_id	Sentence	Label	category
322_sc_1	EN_295	I love how my boss just took that project I have been working hard on away from me for no good reason!	figurative	sarcasm
	KO_295	우리 사장님이 아무 이유 없이 제가 열심히 해온 프로젝트를 뺏어가셨어요. 정말 감동적이에요!	figurative	sarcasm
	ZH_295	老板毫无理由地把我辛苦做的项目拿走了，我真是太爱他了。	figurative	sarcasm

Table 4: Example of FLUTE-bi parallel dataset. EN-KO-ZH sentences which share same ‘Parallel_id’ have same meaning, label and category.

Parallel_id	Lang_id	dialogue	Label	category
2295_m_1	EN_1200	Motive: The car pummeled the toy. A: Did you see what happened in the street just now? B: Yes, it was unbelievable! The car _____ the toy right over. (...) f 1. caressed 2. hummed 3. patted 4. pummeled (answer: 4)	figurative	metaphor
	KO_1200	A: What happened in the car park yesterday? B: A car hit my bike at _____. (…)	figurative	metaphor
	ZH_1200	A: How bad was that accident? B: The car became an easily crushed _____ , completely deformed. (…)	figurative	metaphor

Table 5: Example of FLUID parallel dataset. The dialogues of KO and ZH are translated into English for understanding, while the actual data is in Korean and Chinese. Answer choices are only shown in English, while Korean and Chinese are omitted for space reasons.

categories (idiom, metaphor, simile, and sarcasm) yielding 2,400 parallel instances (Table 4). By embedding figurative choices in realistic dialogue, FLUID QA[1] provides a scalable framework for evaluating usage-level figurative competence in multilingual contexts.

3.3 FLUTE-bi: A Diagnostic Baseline for Figurative Recognition

To aid interpretation of FLUID QA results, we introduce FLUTE-bi, a sentence-level classification task that isolates recognition ability from pragmatic usage. Adapted from the original FLUTE dataset, which used paired sentences for NLI-style inference, we reformulate it as a single-sentence binary classification task to prevent models from exploiting paired cues.

The dataset contains 800 sentence pairs parallel to three languages (Table 5), each with a figurative and a literal version. Each sentence is labeled as figurative or literal. FLUTE-bi provides a reference point for baseline ‘recognition’ without discourse context.

4 Experiments

To assess how large language models (LLMs) handle figurative language at both recognition and usage levels, we evaluate them on two tasks:

- **FLUTE-bi**, which tests semantic recognition via binary classification.
- **FLUID QA**, which probes pragmatic usage through dialogue-based figurative expression selection.

This dual-task evaluation allows us to disentangle semantic understanding from discourse-level application and to reveal how models perform across languages, rhetorical categories, and model types.

4.1 Models

We evaluate a diverse pool of LLMs that vary in architecture, training scale, and degree of language specialization. The selected models fall into three categories.

(1) Universal proprietary models, such as Claude 3.5 Sonnet (Anthropic, 2024) and Gemini 2.0 Flash

[1] The datasets are publicly available at https://github.com/beammeup1229/FLUID_QA

(Team Gemini, 2023), are widely trained on multilingual corpora and optimized for broad language understanding. These models are included as strong multilingual baselines, expected to perform robustly across English, Korean, and Chinese.

(2) Locally specialized proprietary models including Yiyan (Yu et al., 2021; ZH) and HyperClova (Yoo et al, 2024; KO), are pre-trained or fine-tuned to perform well in their respective target languages. Their inclusion allows us to assess the impact of language-specific adaptation on figurative competence.

(3) Open-source models include both base and language-adapted configurations. All open-source models are size-matched at approximately 7–8 billion parameters to control for scale variation. We use LLaMA 3.1-EN-8B (Dubey et al., 2024), the English base model released by Meta, alongside community fine-tuned variants, LLaMA 3.1-KO-8B and LLaMA 3.1-ZH-8B, which were independently adapted by third-party developers using Korean and Chinese corpora, respectively. We also include Qwen 2.5-7B (Yang et al., 2024;ZH) and Exaone 3.5-7.8B (An et al., 2024; KO), two open-source models that were pretrained by large technology firms in their respective language regions. Compared to community fine-tuned versions of LLaMA 3.1, these models were developed using proprietary infrastructure and large-scale in-house resources, allowing more control over pretraining data and objectives.

We intentionally exclude the GPT family (e.g., GPT-4, GPT-4o) from evaluation, as GPT-4o was involved in data generation and may have partial exposure to test content. To verify the risk of contamination, we conducted control experiments detailed in Appendix D. Full model version details are provided in Appendix E.

4.2 Task Setup

Each task is evaluated under distinct conditions to capture different dimensions of figurative competence. FLUID QA is conducted in a zero-shot setting, where models have access to the full multi-turn dialogue but receive no in-context examples. Each item ends with a cloze-style prompt, requiring the model to select the most pragmatically appropriate figurative expression from four candidates. This setup isolates discourse-level reasoning by eliminating external cues and emphasizing context-sensitive interpretation within the dialogue itself.

FLUTE-bi is tested under 0-shot, 5-shot, and 10-shot conditions. Each model performs binary classification on individual sentences, determining whether the expression is figurative or literal. This setting allows us to examine in-context learning effects on basic recognition, independent of dialogue context.

We report macro F1 scores for both benchmarks, broken down by language (EN, ZH, KO) and model type (proprietary vs. open-source). This allows us to examine cross-linguistic consistency, model-specific sensitivity, and the extent to which performance on recognition tasks correlates with usage-level competence.

All experiments were conducted under a single A100 using fixed seeds and consistent formatting. For open-source models, we used official Hugging Face checkpoints. Proprietary models were accessed via public APIs. For few-shot settings, examples are drawn randomly but constrained to maintain category balance.

5 Results

This section presents the performance of large language models (LLMs) on FLUTE-bi for sentence-level recognition and FLUID QA for discourse-level contextual usage. Results are reported by task, language (EN, KO, ZH), model type (proprietary vs. open-source), and few-shot conditions where applicable. The findings reveal consistent dissociations between recognition and usage, substantial disparities across languages, and category-sensitive vulnerabilities in figurative understanding.

5.1 Sentence-Level Figurative Recognition (Binary Classification)

As shown in Table 6, proprietary models consistently outperform open-source models across all languages and conditions. Claude 3.5 maintains scores above 0.84 in English under all shot settings, and across the FLUTE-bi task English and Chinese show broadly comparable performance, whereas Korean consistently yields lower results than the other two languages.

Language-specified models largely follow this dominance pattern, though Exaone represents a notable exception by showing gradual improvements in Korean performance as the

F1-score	0-shot			5-shot			10-shot		
	EN	KO	ZH	EN	KO	ZH	EN	KO	ZH
Claude 3.5 Sonnet	0.84	0.79	0.85	0.84	0.79	0.83	0.84	0.79	0.83
Gemini 2.0 Flash	0.80	0.70	0.73	0.75	0.71	0.76	0.82	0.74	0.81
Yiyan	0.82	0.78	0.83	0.81	0.76	0.83	0.82	0.76	0.81
HyperClova	0.68	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
Qwen2.5-7B	0.76	0.70	0.75	0.76	0.69	0.76	0.75	0.68	0.73
Exaone3.5-7.8B	0.76	0.67	0.74	0.74	0.73	0.66	0.75	0.74	0.69
Llama3.1-EN-8B	0.70	0.58	0.72	0.70	0.65	0.71	0.69	0.68	0.70
Llama3.1-KO-8B	0.60	0.49	0.50	0.64	0.62	0.62	0.64	0.63	0.68
Llama3.1-ZH-8B	0.64	0.62	0.68	0.68	0.68	0.73	0.72	0.58	0.76

Table 6: Binary classification performance (macro F1) on FLUTE-bi across zero-, five-, and ten-shot settings. Results are grouped by model type and language. Cell shading reflects F1 score magnitude, with the brightest color for the lowest score and the darkest for the highest.

F1-score	EN	KO	ZH
Claude 3.5 Sonnet	0.72	0.70	0.75
Gemini 2.0 Flash	0.75	0.71	0.79
Yiyan	0.72	0.63	0.77
HyperClova	0.58	0.52	0.21
Qwen2.5-7B	0.51	0.41	0.46
Exaone3.5-7.8B	0.59	0.52	0.31
LLaMA 3.1-EN-8B	0.60	0.42	0.20
LLaMA 3.1-KO-8B	0.57	0.40	0.34
LLaMA 3.1-ZH-8B	0.50	0.39	0.50

Table 7: Figurative expression selection performance (F1) on FLUID QA by model and language. Performance reflects discourse-level appropriateness under zero-shot conditions. Cell shading reflects F1 score magnitude, with the brightest color for the lowest score and the darkest for the highest.

number of shots increases. Similarly, community fine-tuned models such as LLaMA-KO and LLaMA-ZH achieve partial gains in their respective target languages. Nevertheless, the overall level of open-source models remains below that of proprietary baselines. HyperClova, despite being specialized for Korean, falls short of expectations, a result that appears to stem from its limited capacity for figurative language processing and broader generalization.

Few-shot prompting shows particularly strong effects in low-baseline languages such as Korean and Chinese, as well as in language-specified models (e.g., Gemini, Exaone, LLaMA-KO, LLaMA-ZH). In contrast, English sees only marginal gains due to its already high baseline. This suggests that in-context learning functions as a compensatory signal in restricted-resource languages but provides limited additional benefit in high-performing languages.

In sum, sentence-level recognition is relatively tractable and can be supplemented through few-shot learning. However, performance remains strongly constrained by the proprietary–open-source divide, the persistent weakness of Korean compared to English and Chinese, and the limited effectiveness of language specialization.

5.2 Figurative Expression Selection in Dialogue (FLUID QA)

Table 7 presents results for FLUID QA, which evaluates figurative language usage in dialogue contexts. Compared to the sentence-level recognition results in Table 6, performance drops sharply across all models, underscoring the greater difficulty of discourse-level reasoning. Proprietary models again lead, with Claude 3.5 and Gemini achieving the strongest results across languages. However, the performance gap between proprietary and open-source models is even wider than in recognition, as open-source systems struggle to generalize from sentence-level understanding to dialogue usage.

In terms of language dominance, the balance between English and Chinese observed in FLUTE-bi does not hold. English remains relatively stable across most models, while Chinese shows clear under-generalization and records the lowest scores. Korean outperforms Chinese but remains below English, forming a hierarchy of EN > KO > ZH at the usage level.

Language-specified models consistently follow the order English > target language > non-target language except Chinese-specified models such as Yiyan achieves its highest score in Chinese, followed by English, and lowest in Korean. LLaMA-ZH reports same performance in English and Chinese. Conversely, Korean-specialized models such as HyperClova and Exaone performs best in English, second in Korean, and worst in Chinese. Community fine-tuned models like LLaMA-KO show the same tendency. Thus, they demonstrate some relative advantage in their target languages over non-target ones, but often fail to surpass English and remain far behind proprietary baselines.

Overall, FLUID QA results reveal that figurative usage in dialogue is substantially harder than recognition. The performance gap between proprietary and open-source models widens, and cross-lingual disparities intensify: English remains strongest, while Chinese proves most vulnerable, and Korean continues to lag behind English. These findings highlight the structural challenge of discourse-level pragmatic reasoning in multilingual figurative contexts, showing that language specialization provides partial benefits but not decisive advantages at the usage level.

5.3 Summary of General Trends

Several generalizable trends emerge from these findings:

Recognition–Usage Gap Across all models, performance on FLUID QA drops substantially compared to FLUTE-bi, confirming that discourse-level pragmatic usage is considerably more challenging than sentence-level recognition. This recognition–usage gap underscores that figurative competence cannot be reduced to lexical or surface-level processing alone.

Language Dominance Patterns of language dominance shift between tasks. In FLUTE-bi, English and Chinese are broadly comparable while Korean lags behind, but in FLUID QA, English emerges as the clear leader, Korean moves to a middle position, and Chinese becomes the weakest. This indicates that discourse-level figurative reasoning amplifies cross-lingual disparities and exposes vulnerabilities that are not apparent at the recognition level.

Language Specialization Language-specified shows some relative advantage for the target language over unrelated ones, but specialization does not translate into decisive gains: target-language scores do not always surpass of English, and proprietary multilingual models remain dominant. Thus, specialization offers partial alignment but limited practical benefit for discourse-level usage.

Proprietary vs. Open-Source Proprietary models maintain a clear advantage across both tasks, but the gap widens in usage. Open-source systems, including community fine-tuned variants, struggle to generalize from recognition to usage, revealing the difficulty of transferring surface-level competence to discourse-level reasoning.

In-Context Learning Few-shot prompting improves recognition performance in low-baseline languages such as Korean and Chinese, confirming its compensatory role in FLUTE-bi. Yet in FLUID QA, even full dialogue context fails to yield comparable benefits, suggesting that natural conversational input does not substitute for effective supervision in pragmatic reasoning.

Together, these findings indicate that figurative competence in LLMs is multi-layered and fragile. While recognition can be boosted with in-context learning, usage in dialogue remains structurally difficult, shaped by entrenched English dominance, uneven cross-lingual generalization, and the limited effectiveness of language specialization. This reinforces the importance of evaluating pragmatic reasoning and discourse fit beyond traditional classification tasks.

6 Category-Level Analysis

While Section 5 demonstrates that LLMs struggle to apply figurative language in dialogue, it remains unclear whether this difficulty is uniform across rhetorical categories.

F1-score	EN				KO				ZH			
	SM	ME	ID	SC	SM	ME	ID	SC	SM	ME	ID	SC
Claude 3.5 Sonnet	0.79	0.73	0.90	0.39	0.84	0.57	0.90	0.45	0.92	0.65	0.90	0.50
Gemini 2.0 Flash	0.82	0.78	0.91	0.41	0.85	0.53	0.95	0.45	0.64	0.82	0.83	0.41
Yiyan	0.84	0.80	0.93	0.23	0.84	0.55	0.82	0.28	0.65	0.86	0.94	0.20
HyperClova	0.65	0.69	0.65	0.25	0.51	0.49	0.83	0.20	0.23	0.25	0.18	0.16
Qwen2.5-7B	0.58	0.57	0.65	0.18	0.54	0.37	0.54	0.16	0.68	0.64	0.59	0.12
Exaone3.5-7.8B	0.65	0.78	0.70	0.18	0.66	0.45	0.78	0.18	0.44	0.39	0.35	0.12
Llama3.1-EN-8B	0.67	0.78	0.74	0.16	0.54	0.37	0.54	0.19	0.21	0.19	0.20	0.16
Llama3.1-KO-8B	0.25	0.18	0.30	0.20	0.44	0.41	0.51	0.23	0.32	0.37	0.47	0.18
Llama3.1-ZH-8B	0.58	0.58	0.67	0.16	0.58	0.57	0.66	0.18	0.56	0.54	0.65	0.21

Table 8: FLUID QA Category-Level F1. This table shows model performance broken down by rhetorical category —simile (SM), metaphor (ME), idiom (ID), and sarcasm (SC) — highlighting how figurative usage difficulty varies across languages and categories. Cell shading reflects F1 score magnitude, with the brightest color for the lowest score and the darkest for the highest.

Gold label	Predicted
simile	metaphor
metaphor	simile
idiom	metaphor
sarcasm	idiom

Table 9: Confusion trends in FLUID QA errors. Each row shows frequent mismatches between gold and predicted rhetorical types, revealing systematic substitution patterns.

To refine our understanding of usage-level competence, we break down performance by rhetorical category (Section 6.1) and examine the systematic confusion patterns (Section 6.2) that emerge when models fail on FLUID QA.

6.1 Performance by Category

Table 8 presents average F1 scores across four rhetorical types (idiom, metaphor, simile, and sarcasm) within the FLUID QA task. The results reveal striking asymmetries in model performance.

Sarcasm emerges as the most difficult type across all models and languages, with F1 scores consistently falling below 0.25, which is close to random guessing among four options. This likely reflects models' limited ability to detect ironic stance or contradiction in pragmatic context, an inference that requires recognizing tone and social intent rather than just semantic similarity.

Idioms, by contrast, are consistently the easiest category, where fixed syntactic forms likely aid pattern recognition.

Metaphors and similes show moderate and unstable performance. While similes sometimes benefit from surface cues (e.g., "like," "as"), metaphors require more abstract conceptual mapping, leading to model confusion, especially when figurative interpretation depends on broader discourse coherence.

These results confirm that figurative usage difficulty is not monolithic: each category poses distinct pragmatic demands, and current models handle them with uneven reliability. Notably, higher performance on idioms suggests that LLMs can succeed when strong lexical and syntactic signals are available, while low performance on sarcasm and metaphor reflects the absence of discourse-level abstraction and pragmatic calibration.

6.2 Systematic Figurative Confusion in Failed QA Judgments

To further understand the nature of usage-level failures, we analyze the rhetorical types of expressions that models selected as incorrect answers. While FLUID QA does not require rhetorical classification at inference time, we retroactively map predictions to categories and compute category-level confusion trends. This allows us to examine whether specific types of figurative expressions are systematically confused in usage-level errors. Across languages and models, frequent confusion patterns emerge as shown in Table 9:

Similes are frequently confused with metaphors, and vice versa. This suggests models rely on surface analogical similarity rather than discourse function.

Idioms are often replaced by metaphors, especially when their usage requires contextual grounding rather than lexical familiarity.

Sarcasm is routinely misinterpreted as idiomatic or literal, reflecting models' difficulty with implicit stance recognition and affective pragmatics.

These patterns suggest that usage-level errors are not random. Instead, they reflect internal biases: when pragmatic reasoning fails, models fall back on semantically similar or syntactically familiar expressions, even when inappropriate in discourse.

In sum, usage-level failure is category-sensitive and structurally patterned. By tracing how and why specific types of figurative meaning break down, especially under an ambiguous context, FLUID QA enables deeper diagnosis of the boundaries of discourse-level reasoning in LLMs.

7 Conclusion

Our results reveal a consistent and substantial gap between recognition and usage: while many models achieve high scores on FLUTE-bi, their performance drops markedly on FLUID QA. This divergence highlights fundamental limitations in discourse-level pragmatic reasoning, especially when tasks demand sensitivity to stance, irony, or nuanced conversational intent. Even multilingual and language-specialized models struggle to generalize their recognition capabilities to dialogue settings, suggesting that lexical familiarity alone is insufficient for discourse-level figurative competence.

Moreover, category-level analysis highlights that figurative language usage difficulty is not uniform. Idioms benefit from structural regularity, whereas sarcasm and metaphor expose deeper weaknesses in context modeling. When models made errors, their choices were not random. Rather, they reflect systematic rhetorical confusion, such as substituting a metaphor for a simile or misreading sarcasm as a literal statement. These patterns suggest fallback behavior based on semantic or syntactic proximity, rather than context-sensitive reasoning.

By reframing figurative language competence as a usage-level, dialogue-grounded ability, FLUID QA offers a new lens for diagnosing communicative reasoning in LLMs. It fills a critical gap between isolated recognition and real-world interaction, and establishes a scalable framework for evaluating the pragmatic fluency of multilingual systems. As figurative language is pervasive and socially loaded, future models must learn not only to detect it, but to use it appropriately, reflecting speaker goals, emotional tone, and cultural context. Our benchmark provides the foundation for that next step.

8 Limitations

Our benchmark focuses exclusively on four rhetorical categories and three languages, limiting its generalizability to broader figurative phenomena like humor. While we include a range of proprietary and open-source models, our evaluation does not systematically vary model size, leaving open questions about the relationship between scale and figurative competence. Furthermore, our evaluation is limited to zero- and few-shot settings, and does not explore fine-tuning or instruction-tuning effects. Future work should expand the category space, incorporate cultural variation more systematically, and investigate adaptive methods for pragmatic alignment.

Acknowledgments

We sincerely thank the anonymous reviewers for their constructive feedback and thoughtful suggestions, which have greatly helped improve the quality of this work. We also acknowledge that this work was supported by the Yonsei University Research Fund of 2024 (2024-22-0317).

References

Anthony M. Paul. (1970). Figurative language. Philosophy & Rhetoric, 225-248.

Gibbs Jr, R. W. (1994). Figurative thought and figurative language.

Jerry A. Fodor & Jerrold J. Katz.1964. The structure of language. Englewood Cliffs, N.J.,: Prentice-Hall. Edited by Jerrold J. Katz.

Roberts, Richard M. and Roger Kreuz. 1994. “Why Do People Use Figurative Language?” Psychological Science 5: 159 - 163.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative Language in Recognizing Textual Entailment. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3354–3361, Online. Association for Computational Linguistics.

Yang Liu, Melissa Xiaohui Qin, Hongming Li, Chao Huang. (2024). Revisiting a Pain in the Neck: Semantic Phrase Processing Benchmark for Language Models. arXiv preprint arXiv:2405.02861.

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.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas, G., & Bhatia, A. (2023). Lost in translation: large language models in non-English content analysis. arXiv preprint arXiv:2306.07377.

Dong, Guoliang, Haoyu Wang, Jun Sun and Xinyu Wang. “Evaluating and Mitigating Linguistic Discrimination in Large Language Models. 2024. ArXiv abs/2404.18534.

Emmy Liu, 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, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Huiyuan Lai and Malvina Nissim. 2022. Multi-Figurative Language Generation. In Proceedings of the 29th International Conference on Computational Linguistics, pages 5939–5954, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and Multi-cultural Figurative Language Understanding. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Geetanjali Rakshit and Jeffrey Flanigan. 2022. FigurativeQA: A Test Benchmark for Figurativeness Comprehension for Question Answering. In Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 160–166, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya.

2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. In Findings of the Association for Computational Linguistics: ACL 2024, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Son, G., Yoon, D., Suk, J., Aula-Blasco, J., Aslan, M., Kim, V. T., ... & Kim, S. (2024). MM-Eval: A Multilingual Meta-Evaluation Benchmark for LLM-as-a-Judge and Reward Models. arXiv preprint arXiv:2410.17578.

Masaru Yamada. 2023. Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. In Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Rezaeimanesh, S., Hosseini, F., & Yaghoobzadeh, Y. (2024). A Comparative Study of LLMs, NMT Models, and Their Combination in Persian-English Idiom Translation. arXiv preprint arXiv:2412.09993.

Khoshtab, P., Namazifard, D., Masoudi, M., Akhgary, A., Sani, S. M., & Yaghoobzadeh, Y. (2024). Comparative Study of Multilingual Idioms and Similes in Large Language Models. arXiv preprint arXiv:2410.16461.

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

Gao, Y., Wang, R., & Hou, F. (2024, December). How to design translation prompts for ChatGPT: An empirical study. In Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops (pp. 1-7).

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024. Creative and Context-Aware Translation of East Asian Idioms with GPT-4. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9285–9305, Miami, Florida, USA. Association for Computational Linguistics.

Singh, P., Patidar, M., & Vig, L. (2024). Translating Across Cultures: LLMs for Intralingual Cultural Adaptation. arXiv preprint arXiv:2406.14504.

Sui He. 2024. Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 316–326, Sheffield, UK. European Association for Machine Translation (EAMT).

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual Multi-Figurative Language Detection. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

Simone A Sprenger. 2003. Fixed expressions and the production of idioms. Ph.D. thesis, Radboud University Nijmegen Nijmegen.

Knappe, Gabriele. "Idioms and fixed expressions." English historical linguistics: An international handbook (2012): 177-196.

Hwichan Kim, Jun Suzuki, Tosho Hirasawa, and Mamoru Komachi. 2024. Pruning Multilingual Large Language Models for Multilingual Inference. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9921–9942, Miami, Florida, USA. Association for Computational Linguistics.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

. Bradley, Ralph Allan, and Milton E. Terry.(1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39.3/4: 324-345.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale Cloze Test Dataset Created by Teachers. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

. Anthropic, A. I. "Claude 3.5 sonnet model card addendum." Claude-3.5 Model Card 3.6 (2024).

Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, Haifeng Wang. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137.

Yoo, Kang Min, et al. "Hyperclova x technical report. 2024. arXiv preprint arXiv:2404.01954.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... & Qiu, Z. (2024). Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.

Research, L. G., An, S., Bae, K., Choi, E., Choi, K., Choi, S. J., ... & Yun, H. (2024). EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. arXiv preprint arXiv:2412.04862.

A Cultural Adaptation translation prompt

You are a professional {Korean/Chinese} translator performing a cultural adaptation translation from a foreign culture to {Korean/Chinese} culture. Given an original sentence from English figurative language dataset, your task is translating English sentence to {Korean/Chinese} {figurative Category} sentence by using cultural adaptation strategies down below:

In the translation of Culture-specific items, Eirlys E. Davies.(2003) defines the following translation strategies;

In the translation of Culture-specific items, Davies defines the following translation strategies:

1. Addition is when more information is added simultaneously with the transfer from source culture to target culture, for example: eating at Wendy's → eating at Wendy's, an American international fast food restaurant chain

2. Omission is a strategy when a word or a phrase is omitted from the target culture when no equivalents can be found, for example: getting a taco from taco bell → getting a taco

3. Globalization is a strategy of exchanging cultural elements of the text with more general and neutral words, to match it with the target language culture, for example: Kimono → Traditional garment; Hamburger → Burger; Greek yoghurt → Curd etc.

4. Localization is trying to find an appropriate equivalent of the CSI in the target language, for example, sausage → kebab; mentos → paan; etc.

5. Transformation is an alteration of a CSI to another CSI which is not a local equivalent but an altered/distorted version, familiar to the target language audience, for example: football game → Local cricket match; mentos → namkeen (alteration of CSI); pastry → halwa (no close equivalent so altered the CSI); etc.

Original English sentences: {sentence}

B QA generation prompt

You are a {English/Korean/Chinese} teacher who wants to make a cloze style dialogue QA for figurative language understanding and your task is to create multiple-choice questions that require selecting the appropriate word for a {category} statement.

Each question consists of a prompt and four choices. Follow the given guidelines to generate them.

1. Prompt

The given sentence contains a {category} expression. Construct a three or four-turn dialogue that includes the given sentence.

Ensure consistency in using either formal or informal speech throughout the conversation.

Indicate the {category} part by replacing it with "_______".

2. Answer Choices

Provide one correct answer and three misleading incorrect choices for the blank.

- incorrect choices should contain one antonym/irrelevant word, two synonyms but have different meanings.
- Separate each answer choice with a new line (\n) and numbering it.
- Ensure that the correct answer is not too obvious.
- Also, indicate which choice is the correct answer.

Target Sentence: {FLUTE_figurative sentence}

Paired literal Sentence: {FLUTE_literal sentence}

Explanation of Target Sentence: {explanation}

C Impact of Translation Strategies: Literal vs. Cultural Adaptation

	EN	KO_C	ZH_C	KO_L	ZH_L
Claude	0.84	0.79	0.85	*0.73*	*0.73*
Gemini	0.8	0.7	0.73	*0.56*	*0.6*

Table 10: Impact of cultural adaptation (C) vs. literal translation (L) on FLUTE-bi performance. Scores are measured using F1-score. We consider 'EN' as baseline.

To assess the actual impact of cultural adaptation, we translated the Korean and Chinese data using literal (non-adaptive) translations and evaluated model performance on the FLUTE-bi classification task (F1-score). The experiments were conducted in 0-shot setting with Claude 3.5 Sonnet and Gemini 2.0 Flash, which showed the strongest performance in Table 6.

As shown on Table 10, we observed consistent drops in F1 scores for both languages compared to the culturally adapted versions. These results suggest that literal translations weaken figurative meaning and increase interpretive ambiguity due to translationese effects, thereby hindering model recognition performance.

D Assessing Potential Contamination from GPT-4o

FLUTE-bi	0-shot			5-shot		
	EN	KO	ZH	EN	KO	ZH
Claude	0.84	0.79	0.85	0.84	0.79	0.83
Yiyan	0.82	0.78	0.83	0.81	0.76	0.83
GPT 4o	*0.83*	*0.74*	*0.82*	*0.83*	*0.78*	*0.83*

Table 11. FLUTE-bi binary classification results (macro F1) comparing GPT-4o with non-GPT proprietary models (Claude 3.5, Yiyan). GPT-4o shows comparable performance and follows the same cross-lingual trend suggesting no clear English-specific contamination.

QA	EN	KO	ZH
Claude	0.72	0.70	0.75
Yiyan	0.72	0.63	0.77
GPT 4o	*0.79*	*0.72*	*0.77*

Table 12. FLUID QA figurative usage results (F1) for GPT-4o and comparison models. GPT-4o follows the same overall tendency, though slightly higher English scores were observed. While this does not provide strong evidence of contamination, the possibility cannot be fully ruled out, motivating the exclusion of GPT models from the main evaluation.

Since GPT-4o was used in data generation, a concern is that it might have contaminated the benchmark. If contamination had occurred, GPT-4o would be expected to substantially outperform non-GPT models due to partial exposure to the data. To examine this, we conducted control experiments using models that showed relatively strong performance in each language in the main results (Tables 6 and 7). Specifically, we compared GPT-4o with Claude 3.5 Sonnet and Yiyan, while excluding HyperClova due to its limited generalization capacity.

Our results do not support the contamination scenario. On FLUTE-bi, GPT-4o performed comparably to Claude 3.5 and Yiyan, following the same cross-lingual pattern (Table 11). On FLUID QA, GPT-4o also aligned with the general hierarchy. However, we did observe relatively higher scores for English (EN) in some cases (Table 12). While these gains are more plausibly explained by variability rather than systematic bias, we cannot fully rule out the possibility of contamination.

For this reason, we fully excluded the GPT family from evaluation to eliminate any residual risk of data, evaluation leakage. In addition, all Korean and Chinese drafts were carefully post-edited by native speakers to remove translationese and ensure cultural and pragmatic adequacy.

E Model Details

Open /Closed	Group	Name	Target language	Version (company)	note
Proprietary	Universal	Claude 3.5 Sonnet	multilingual	claude-3-5-sonnet-20241022 (Anthropic)	
		Gemini 2.0 Flash	multilingual	gemini-2.0-flash (Google)	
	Local	Yiyan	Chinese	ernie-4.0-turbo-8k (Baidu)	
		HyperClova	Korean	HCX-DASH-001 (Naver)	
Open-source	Tech-company	Qwen2.5-7B	Chinese	Qwen2.5-7B-Instruct[3] (Alibaba)	Models are from huggingface
		Exaone3.5-7.8B	Korean	EXAONE-3.5-7.8B-Instruct[4] (LG AI)	
		Llama3.1-EN-8B	English	Llama-3.1-8B-Instruct[5] (Meta)	
	community fine-tuned	Llama3.1-KO-8B	Korean	llama3.1_korean_v1.1_sft_by_aidx[6]	
		Llama3.1-ZH-8B	Chinese	Llama3.1-8B-Chinese-Chat[7]	

[3] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/SEOKDONG/llama3.1_korean_v1.1_sft_by_aidx

[7] https://huggingface.co/ shenzhi-wang/Llama3.1-8B-Chinese-Chat