How Order-Sensitive Are LLMs?
OrderProbe for Deterministic Structural Reconstruction

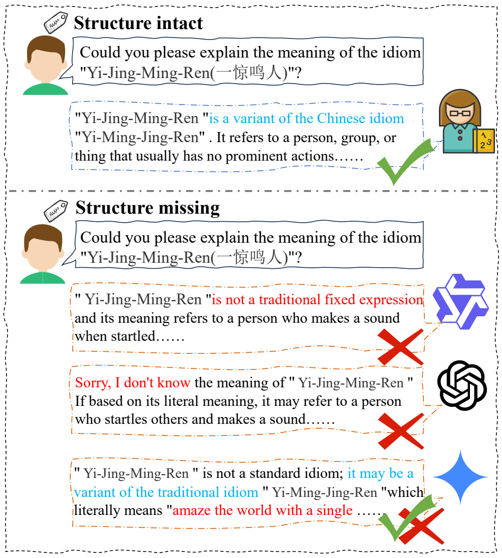
Anonymous ACL submission

Abstract

Large language models (LLMs) excel at semantic understanding, yet their ability to reconstruct internal structure from scrambled inputs remains underexplored. Sentence-level restoration is ill-posed for automated evaluation because multiple valid word orders often exist. We introduce **OrderProbe**, a deterministic benchmark for structural reconstruction using fixed four-character expressions in Chinese, Japanese, and Korean, which have a unique canonical order and thus support exact-match scoring. We further propose a diagnostic framework that evaluates models beyond recovery accuracy, including semantic fidelity, logical validity, consistency, robustness sensitivity, and information density. Experiments on twelve widely used LLMs show that structural reconstruction remains difficult even for frontier systems: zero-shot recovery frequently falls below 35%. We also observe a consistent dissociation between semantic recall and structural planning, suggesting that structural robustness is not an automatic byproduct of semantic competence.

1 Introduction

Large language models (LLMs) have demonstrated strong performance on semantic understanding and generation tasks. However, their robustness to internal structural disruption remains an open question (OpenAI et al., 2024; DeepSeek-AI et al., 2025). In human language processing, listeners can often recover meaning from structurally disordered inputs with little effort. Replicating this evaluation for machine models, however, poses a fundamental methodological challenge. The flexibility of natural language syntax means that a scrambled sentence often admits multiple valid permutations, preventing a unique ground truth and making deterministic evaluation difficult. This ambiguity limits sentence-level restoration benchmarks and often conflates structural ability with general fluency.



Could you please explain the meaning of the idiom "Yi-Jing-Ming-Ren(一惊鸣人)"?

"Yi-Jing-Ming-Ren "is a variant of the Chinese idiom "Yi-Ming-Jing-Ren" . It refers to a person, group, or thing that usually has no prominent actions……

Could you please explain the meaning of the idiom "Yi-Jing-Ming-Ren(一惊鸣人)"?

" Yi-Jing-Ming-Ren "is not a traditional fixed expression and its meaning refers to a person who makes a sound when startled……

Sorry, I don't know the meaning of " Yi-Jing-Ming-Ren " If based on its literal meaning, it may refer to a person who startles others and makes a sound……

" Yi-Jing-Ming-Ren " is not a standard idiom; it may be a variant of the traditional idiom " Yi-Ming-Jing-Ren "which literally means "amaze the world with a single ……

Figure 1: **OrderProbe example illustrating semantic–structure dissociation.** Given a scrambled four-character expression, humans can often recover the unique canonical order using semantic anchors and structural priors. Many LLMs, however, produce correct meanings while failing exact reconstruction, revealing a gap between semantic recall and structural planning.

Existing robustness evaluations largely target surface noise or external instruction perturbations rather than internal ordering constraints(Wang et al., 2021; Ribeiro et al., 2020; Li et al., 2024). Although these benchmarks have advanced robustness evaluation, they do not directly probe whether a model can reconstruct a globally constrained sequence from scrambled components. This omission is particularly important given evidence that Transformer models can exhibit limited sensitivity to word order (Sinha et al., 2021), and may fail to generalize structural relations under reversed or reordered contexts (Berglund et al., 2024).

To isolate structural reconstruction under deterministic scoring, we use fixed four-character expressions in Chinese, Japanese, and Korean as the evaluation unit. Sentence-level restoration is ill-posed because scrambled sentences often admit

Benchmark	**Year**	**Language**	**Target Unit**	**Perturbation Level**	**Core Task**
General Robustness Benchmarks					
RoTBench (Ye et al., 2024)	2024	En	Tool-use I/O	Real-world Noise	Tool-use Robustness (Tool Learning)
Instruct-Robust (Li et al., 2024)	2024	En	Instructions	External Injection	Instruction-Following Robustness
ReDial (Lin et al., 2025)	2025	En (Dialect)	Logic Problems	Dialectal Variation	Reasoning (Math/Logic)
Idiom & Figurative Language Benchmarks					
MMFLD (Lai et al., 2023)	2023	Multi	Fig. Language	*None* (Static)	Detection / Class.
WenMind (Cao et al., 2024)	2024	Zh	Classical Text	*None* (Static)	QA & Generation
Chengyu-Bench (Fu et al., 2025)	2025	Zh	Idioms	*None* (Static)	Understanding & Use
MChIRC (Wang et al., 2025)	2025	Zh	Idiom-in-Context	*None* (Static)	Multimodal RC / Cloze
Idiom-Trans (Donthi et al., 2025)	2025	Multi	Idioms	*None* (Static)	Translation
OrderProbe (Ours)	**2025**	**CJK**	**Idioms**	**Internal Structure**	**Reconstruction**

Table 1: **Positioning of OrderProbe.** Existing robustness benchmarks mainly test surface-level noise or instruction-level perturbations, while idiom benchmarks treat expressions as static retrieval or transformation targets. OrderProbe uniquely evaluates deterministic internal-order reconstruction with exact-match scoring.

multiple valid reorderings, preventing a unique ground truth. In contrast, four-character expressions typically have a rigid internal sequence and a single canonical form, enabling exact-match evaluation while preserving high semantic density in a minimal structure. This setting also differs from existing idiom and figurative language benchmarks, which mainly test retrieval or interpretation under canonical inputs (Zheng et al., 2019; Lai et al., 2023; Fu et al., 2025) or translation and usage (Donthi et al., 2025), rather than reconstruction from scrambled constituents. Classical understanding datasets similarly emphasize semantics without probing structure recovery under internal perturbation (Cao et al., 2024).

Figure 6 illustrates this distinction. When presented with scrambled characters from the idiom *Yi-Ming-Jing-Ren* ("to become famous overnight"), humans can quickly reconstruct the canonical order based on semantic anchors. By contrast, although LLMs can provide accurate definitions when queried directly, they often fail to recover the correct structure when constrained to use the scrambled components. Instead, models tend to process the characters as independent tokens. This behavior reveals a dissociation where models memorize the semantic entry of an expression but fail to reassemble its canonical structure from its parts.

To systematically evaluate this dissociation, we introduce OrderProbe, a benchmark designed to measure structural reconstruction capabilities in Chinese, Japanese, and Korean. The dataset is constructed through a rigorous pipeline involving multi-source acquisition, expert filtering, and hybrid semantic annotation to ensure linguistic precision. It comprises 3,543 curated samples, covering six syntactic categories and multiple script typologies. Beyond dataset construction, we propose a diagnostic evaluation framework that decomposes model behavior into six dimensions, including reconstruction accuracy, semantic fidelity, logical validity, structural consistency, robustness sensitivity, and information density. Unlike approaches that rely solely on coarse accuracy rankings, this framework enables fine-grained analysis of failure modes that separate semantic recall from structural planning.

Our contributions can be summarized as follows.

- We identify the ambiguity of sentence-level restoration and establish fixed four-character expressions as a deterministic proxy for evaluating structural reconstruction.
- We introduce OrderProbe, a multilingual benchmark comprising 3,543 curated samples across six syntactic categories.
- We propose a diagnostic evaluation framework that moves beyond simple scoring to characterize robustness sensitivity, semantic fidelity, and hallucination behavior.
- We evaluate twelve widely used LLMs on OrderProbe and find that exact reconstruction remains challenging, with zero-shot recovery frequently below 35% even for state-of-the-art systems.

2 OrderProbe

OrderProbe is constructed through a controlled pipeline that transforms raw lexical resources into a deterministic benchmark for structural reconstruction. The pipeline consists of four stages: multi-source collection, expert filtering, semantic reference construction, and permutation-based perturbation generation. This section describes each stage and summarizes the resulting dataset.

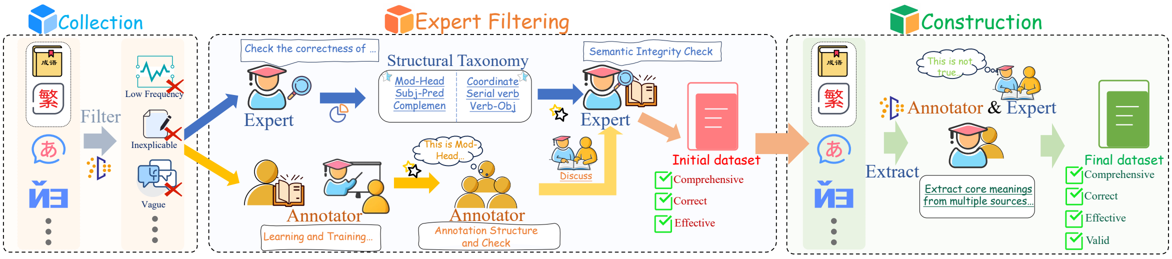


Figure 2: **OrderProbe construction pipeline.** We collect raw four-character expressions from multiple sources, apply expert filtering to obtain canonical roots, build semantic references via dictionary grounding and verified augmentation, and generate the full non-identity permutation space as structurally perturbed evaluation inputs.

2.1 Stage I: Multi-source Collection

We first assemble a candidate pool $\mathcal{C}_{raw}$ by aggregating four-character expressions from multiple publicly available resources, including lexical dictionaries and digital repositories. Each candidate item is associated with a script typology $T \in \{\text{ZH-CN}, \text{ZH-TW}, \text{JA}, \text{KO}\}$. The goal of this stage is to maximize coverage and diversity before quality control. This process yields approximately 4,000 raw candidates.

2.2 Stage II: Expert Filtering

To obtain a linguistically reliable core set, we apply an expert filtering function $\mathcal{F}$ that maps $\mathcal{C}_{raw}$ to a refined set $\mathcal{X}$:

$$\mathcal{X} = \mathcal{F}(\mathcal{C}_{raw}). \tag{1}$$

Filtering is performed by senior linguists and trained annotators with proficiency in the target languages. Items are removed if they are non-standard, ambiguous, modern internet slang, or exhibit multiple competing canonical forms. After cross-lingual alignment and consensus review, this process yields exactly 3,543 canonical expressions in $\mathcal{X}$, which form the benchmark roots.

2.3 Stage III: Semantic Reference Construction

For each canonical expression $x \in \mathcal{X}$, we construct a semantic reference set $\mathcal{S}_x$ to support evaluation beyond exact recovery. We retrieve dictionary definitions d_{dict} as the primary semantic anchor, following common practice in idiom evaluation (Zheng et al., 2019; Fu et al., 2025). To reduce brittleness to surface phrasing, we augment each definition with paraphrastic variants generated by LLMs and verified by annotators, similar to recent benchmark construction protocols (Li et al., 2024). The final reference set is defined as:

$$\mathcal{S}_x = \{d_{dict}\} \cup \{d_{aug}\}. \tag{2}$$

This hybrid construction enables robust semantic and logical evaluation without translating across languages.

2.4 Stage IV: Permutation-based Perturbation Generation

The benchmark evaluates whether models can reconstruct the canonical order of an expression from scrambled constituents. Each canonical expression x is represented as a token sequence (t_1, t_2, t_3, t_4) with a fixed order. We apply a perturbation operator $\mathcal{P}$ that generates all non-identity permutations:

$$\mathcal{P}(x) = \{\pi(x) \mid \pi \in S_4, \pi(x) \neq x\}, \tag{3}$$

where S_4 denotes the permutation group over four elements. This yields 23 perturbed inputs per canonical expression, forming a complete combinatorial space for internal reordering. Unlike surface perturbations such as typos or synonym substitutions (Wang et al., 2021; Ribeiro et al., 2020), this operator preserves all lexical content while disrupting only the internal structure.

2.5 Dataset Statistics

The final dataset $\mathcal{D}$ consists of all perturbed variants generated from the canonical set $\mathcal{X}$. It covers four script settings (Simplified Chinese, Traditional Chinese, Japanese, and Korean) and six syntactic categories. The distribution of script typologies and structural patterns is shown in Figure 3. This design supports controlled analyses of how reconstruction varies across writing systems and syntactic templates. We additionally treat Korean Hangul as a phonogrammatic negative control: unlike logographic scripts, scrambling Hangul characters disrupts syllable composition and removes local semantic anchors, making deterministic reconstruction fundamentally harder.

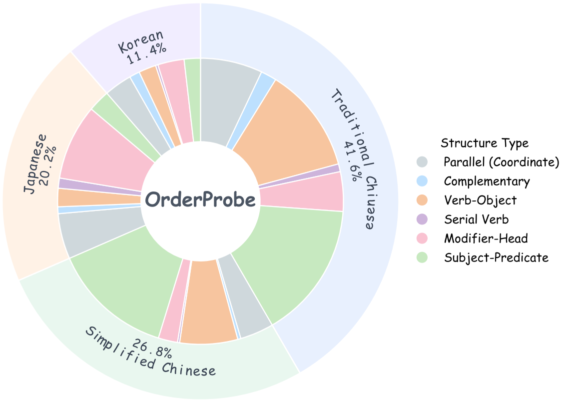


Figure 3: **Overview of OrderProbe.** The chart displays the distribution of script typologies and syntactic structures across the 3543 samples.

Category	Metric	Linguistic Definition & Rationale
Primary	**Recovery Rate**	Exact match of canonical order x. Measures global structural integrity.
Diagnostics	S_{Rob}	Robustness score based on degradation under S_n permutations (MDR/MDA).
	S_{Acc}^{mean}	Semantic fidelity via metaphorical alignment using Cross-Encoders and BERTScore.
	S_{Logic}	NLI-based entailment check between model output e and reference meaning.
	S_{Cons}	Stability across permutation variants, capturing invariance to internal shuffles.
	S_{Info}	Information density penalizing verbosity while rewarding high information-to-token ratio.

Table 2: **Evaluation Metrics on OrderProbe.** We prioritize Recovery Rate as the global structural indicator. Other diagnostic metrics characterize robustness, semantic fidelity, logical validity, invariance, and information density.

3 Evaluation Framework

Beyond deterministic exact-match reconstruction, OrderProbe contributes a multi-dimensional diagnostic evaluation framework that disentangles structural planning from semantic recall. A single aggregate score is insufficient: models may fail recovery due to hallucinated retrieval, positional sensitivity, format mimicry, or verbose “knowledge dumping,” all of which can mask the underlying failure mechanism. We therefore design six complementary metrics(Table 2) computed over the full permutation space, enabling fine-grained failure-mode analysis.

Evaluation Scope. For each canonical expression x, we evaluate model behavior under all 23 non-identity permutations $\mathcal{P}(x)$, yielding $3543 \times 23 = 81{,}489$ perturbed inputs. Each output is scored by exact reconstruction, semantic alignment, and logical entailment, and then aggregated across permutations to quantify sensitivity and invariance. This produces a six-dimensional diagnostic signature per model, rather than only a leaderboard rank.

3.1 Primary Metric: Recovery Rate

The Recovery Rate is the global indicator of structural integrity:

$$\text{Recovery} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{x}_i = x_i). \tag{4}$$

Unlike sentence restoration (ill-posed due to multiple valid reorderings), four-character expressions have a unique canonical form, enabling deterministic evaluation.

3.2 Diagnostic Metrics (Beyond Recovery)

We introduce five diagnostic dimensions to explain *why* recovery fails.

Semantic Fidelity (S_{Acc}^{mean}). We evaluate explanation quality using a tiered hybrid semantic metric integrating cross-encoder relevance, multilingual embedding similarity, and lexical safeguards against entity-level hallucination (Appendix A.2).

Logical Validity (S_{Logic}). To detect fluent but contradictory definitions, we compute entailment probability between generated and reference meanings using a multilingual NLI model:

$$S_{\text{Logic}} = P_{\text{NLI}}(e \Rightarrow r). \tag{5}$$

Structural Consistency (S_{Cons}). A key novelty is a permutation-level invariance score. For each permutation, we measure the gap between the model’s potential best semantic score and its realized mean score, and aggregate both average deviation and worst-case sensitivity across $\mathcal{P}(x)$. We define:

$$S_{\text{Cons}} = (1 - E_{\text{perf}}) \cdot (1 - R_{\text{sens}}), \tag{6}$$

which penalizes both global instability and localized brittleness (Appendix A.3).

Robustness (S_{Rob}). We quantify robustness as degradation relative to the canonical baseline and decompose it into sequential robustness (permutation sensitivity) and structural robustness (variance across syntactic categories), combined via harmonic mean (Appendix A.5).

Information Density (S_{Info}). To penalize verbosity-based metric inflation, we introduce an information density score using a brevity penalty that rewards informative explanations per token (Appendix A.7).

Model Family	Setting	Diagnostic Component Metrics					Main Metric
		S_{Acc}^{mean}	S_{Cons}	S_{Logic}	S_{Rob}	S_{Info}	Recovery. (%)
Tier 1: High-Performance Reasoning Models							
Qwen-3-14B	Zero-shot	0.493	0.591	0.672	0.621	**0.579**	31.062
	+CoT	0.429	0.741	0.811	0.603	0.536	31.527 (↑0.465)
Qwen-3-VL$_{\text{Think}}$	Zero-shot	**0.520**	0.539	0.712	0.645	0.520	30.928
	+CoT	0.471	0.716	0.888	0.632	0.533	48.376 (↑17.448)
Gemini-2.5-Flash	Zero-shot	0.493	0.621	0.646	0.631	0.540	30.487
	+CoT	0.455	0.738	0.738	**0.647**	0.246	41.559 (↑11.072)
Tier 2: Mid-Range General Models							
GPT-4o	Zero-shot	0.463	0.663	0.591	0.616	0.489	23.254
	+CoT	0.431	0.769	0.774	0.622	0.400	37.687 (↑14.433)
GLM-4V$_{\text{Think}}$	Zero-shot	0.446	0.565	0.629	0.591	0.446	21.636
	+CoT	0.413	0.785	0.756	0.594	0.361	30.079 (↑8.443)
DeepSeek-V3.2	Zero-shot	0.472	0.593	0.575	0.609	0.444	20.104
	+CoT	0.482	0.727	**0.948**	0.643	**0.689**	**62.238** (↑42.134)
Tier 3: Efficient & Baseline Models							
DeepSeek-R1	Zero-shot	0.446	0.682	0.465	0.593	0.469	18.596
	+CoT	0.420	0.790	0.739	0.612	0.470	27.729 (↑9.133)
Claude-HaiKu-4.5	Zero-shot	0.448	0.656	0.494	0.612	0.466	18.017
	+CoT	0.398	0.714	0.652	0.551	0.293	21.646 (↑3.629)
Gemma-3-4B	Zero-shot	0.439	0.654	0.578	0.610	0.291	17.094
	+CoT	0.402	0.783	0.655	0.600	0.104	18.027 (↑0.933)
Llama-3-8B	Zero-shot	0.396	0.706	0.399	0.590	0.304	12.167
	+CoT	0.385	**0.815**	0.641	0.586	0.226	24.428 (↑12.261)
Llama-2-7B	Zero-shot	0.411	0.731	0.502	0.588	0.182	12.576
	+CoT	0.412	0.744	0.469	0.602	0.258	18.747 (↑6.171)
Qwen-3-8B	Zero-shot	0.446	0.617	0.487	0.604	0.403	15.786
	+CoT	0.392	0.715	0.492	0.580	0.269	12.594 (↓3.192)

Table 3: **Comparative Assessment of Zero Shot and Chain of Thought Prompting across 12 LLMs on OrderProbe.** This table reports performance metrics encompassing semantic accuracy, consistency and structural recovery.

Takeaway. Together, these metrics expose systematic dissociations such as high semantic fidelity but low structural recovery, or high consistency due to rigid hallucination, enabling diagnosis beyond surface accuracy rankings. Full formulations and implementation details are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

We evaluate all models on the full OrderProbe benchmark containing 3,543 canonical four-character expressions across four script settings (ZH-CN, ZH-TW, JA, KO). Given each scrambled input, models are required to output *(i)* the recovered canonical expression (a single four-character sequence) and *(ii)* a one-line semantic explanation in the corresponding language. All outputs are evaluated by exact-match reconstruction and our diagnostic metrics described in Section 3.

Models. We evaluate twelve widely used LLMs spanning open- and closed-source families. Open-source models include instruction-tuned checkpoints from Qwen (Qwen et al., 2025), DeepSeek (DeepSeek-AI et al., 2025), Llama (Grattafiori et al., 2024), and Gemma, run locally using official weights. Closed-source models include GPT (OpenAI et al., 2024), Gemini, Claude (Anthropic, 2024), and GLM (Team et al., 2025), accessed via official APIs with provider-default settings during Oct 2025–Jan 2026. All checkpoint and deployment details are provided in Appendix A.10.

Prompting Strategy and Output Constraints. We use a unified two-message prompting format (system + user) with strict schema enforcement to minimize format drift. We adopt language-matched prompting: each script setting is evaluated using prompts written in the same language as the input, and models are instructed to produce both the

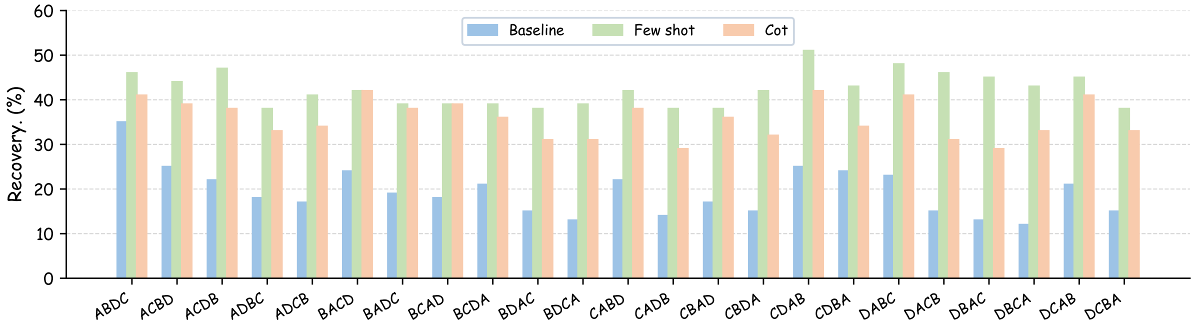


Figure 4: **Performance Comparison across 23 Permutation Patterns.** The chart illustrates recovery rates under baseline, few shot, and chain of thought settings, revealing the impact of anchor displacement on model stability.

Language	**Component Metrics**					**Main Metric**
	S_{Acc}^{mean}	S_{Cons}	S_{Logic}	S_{Rob}	S_{Info}	**Recov. (%)**
Simplified Chinese	**0.4817**	0.6445	**0.6854**	**0.6193**	0.4631	**25.20**
Traditional Chinese	0.4715	0.6184	0.6663	0.5998	0.4040	**20.32**
Japanese	0.4602	0.5897	0.5278	0.6048	**0.5388**	**19.66**
Korean	0.3990	**0.7625**	0.3014	0.5857	0.3906	**5.72**

Table 4: **Comparison across script settings.** Logographic scripts (Chinese/Japanese) yield substantially higher reconstruction accuracy than Korean Hangul, which serves as a phonogrammatic control condition. Component metrics are reported for diagnostic analysis alongside the primary Recovery score.

recovered expression and explanation in that language. We compare three settings: **(i)** Zero-shot (one-line output), **(ii)** Chain-of-Thought (CoT; exactly two lines: reasoning + final answer), and **(iii)** three-shot in-context learning (few-shot; two-line output following exemplars). All prompt templates are provided in Appendix C.

4.2 Overall Results and Analysis

Table 3 summarizes the benchmark results across twelve models under both Zero-shot and CoT settings. Across model families, we observe a consistent semantic–structure dissociation: models can often generate fluent, meaning-aligned explanations while failing to reconstruct the canonical order under permutation noise.

Reconstruction remains challenging. Exact reconstruction is difficult even for strong systems. Under the zero-shot setting, recovery frequently remains below 35%, indicating that structural recovery is not an automatic byproduct of semantic competence. Instead, models tend to behave as if scrambled characters are independent tokens, producing semantically plausible but structurally incorrect outputs.

CoT helps, but gains vary widely. CoT prompting improves recovery for many models, suggesting that stepwise reasoning encourages explicit constraint checking before generation. However, improvements are highly model-dependent: some systems benefit substantially, while others show marginal gains or even regressions due to format drift and unstable decoding.

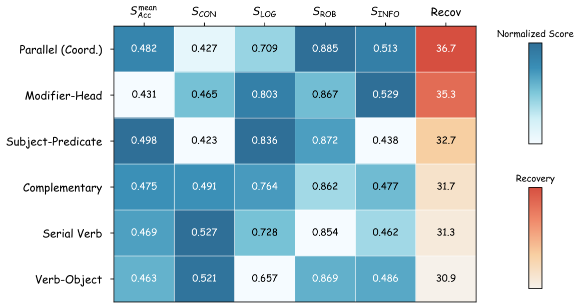
Consistency is not correctness. High structural consistency can occur even when recovery is low. In several cases, models respond with stable but generic explanations across permutations, producing high invariance scores without solving the reconstruction problem. This highlights the necessity of interpreting diagnostic metrics jointly rather than relying on a single dimension.

Summary. Overall, OrderProbe reveals that deterministic structural reconstruction remains a substantial capability gap. Models can retrieve meanings or provide plausible definitions, but often fail to enforce global ordering constraints required for exact canonical reconstruction.

4.3 Permutation Sensitivity and Prompting Effects

Figure 4 compares recovery trends across all 23 permutation configurations under baseline, CoT, and few-shot settings. We observe strong dependence on anchor displacement: permutations that preserve semantically informative characters near their canonical neighborhoods yield higher recovery, while dispersed anchors significantly degrade performance.

Few-shot prompting yields the most uniform improvements across permutations, suggesting that explicit exemplars constrain decoding more effectively than generic reasoning traces. In contrast, CoT gains are less stable and may amplify format drift in weaker models.



	S_{Acc}^{mean}	S_{CON}	S_{LOG}	S_{ROB}	S_{INFO}	Recov
Parallel (Coord.)	0.482	0.427	0.709	0.885	0.513	36.7
Modifier-Head	0.431	0.465	0.803	0.867	0.529	35.3
Subject-Predicate	0.498	0.423	0.836	0.872	0.438	32.7
Complementary	0.475	0.491	0.764	0.862	0.477	31.7
Serial Verb	0.469	0.527	0.728	0.854	0.462	31.3
Verb-Object	0.463	0.521	0.657	0.869	0.486	30.9

Figure 5: **Performance by syntactic category.** Parallel structures are easiest due to redundancy and symmetry cues, while asymmetric dependency patterns remain harder for reconstruction. Metrics are reported to diagnose whether failures arise from semantic gaps, instability, or structural sensitivity.

4.4 Impact of Script Typology

Table 4 compares performance across the four script settings. We observe that reconstruction difficulty is strongly shaped by script typology.

Logographic scripts provide stronger local anchors. Chinese and Japanese exhibit substantially higher recovery than Korean. Logographic characters preserve standalone semantic cues under scrambling, providing local anchors that facilitate global reassembly.

Hangul scrambling is a strict negative control. Korean Hangul behaves differently: internal scrambling destroys syllable composition and removes meaningful local cues. As a result, recovery collapses sharply, confirming that reconstruction depends not only on positional decoding but also on the preservation of interpretable sub-units.

Consistency inflation under extreme ambiguity. Notably, Korean yields high consistency despite low recovery. Under severe uncertainty, models often default to generic, stable explanations across permutations. This yields artificially high invariance but does not reflect true reconstruction ability, reinforcing the need to interpret diagnostic metrics jointly.

4.5 Analysis of Syntactic Patterns

Figure 5 reports results across six syntactic categories, revealing systematic structural variation beyond surface shuffling.

Parallel structures are easiest. Parallel and coordinate patterns are generally easier due to redundancy and symmetric templates. These cues effectively reduce the search space and provide ordering heuristics.

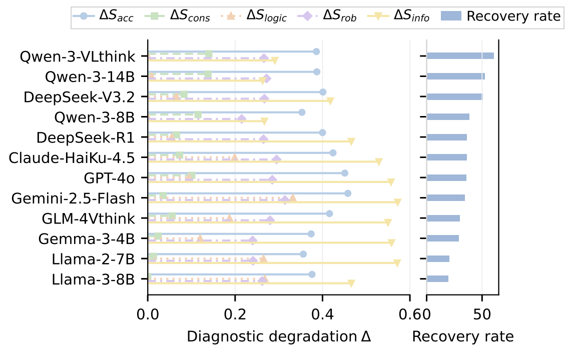


Figure 6: **Degradation under perturbed 3-shot inputs.** Left: per-metric drops ΔS_{acc}, ΔS_{cons}, ΔS_{logic}, ΔS_{rob}, ΔS_{info}, computed as *unperturbed* minus *perturbed 3-shot* under matched inference settings. Right: recovery rate for perturbed 3-shot inputs. Models are ordered by recovery rate (higher is more robust).

Meaning does not imply correct ordering. For subject–predicate patterns, models often recover the underlying proposition but fail to linearize characters into the precise canonical order, confirming that semantic grasp does not guarantee structural planning.

Asymmetric dependencies remain difficult. Verb–object patterns are harder because they rely on directional dependencies that are disrupted by permutation noise, leading to semantically plausible but structurally incorrect guesses.

4.6 Analysis of In-Context Learning

Figure 6 reports results under three-shot prompting.

Perturbations reveal structured degradation rather than uniform noise. Across models, diagnostic drops concentrate on a subset of components instead of moving in lockstep, forming consistent “degradation fingerprints” per model. This suggests perturbations disrupt specific capabilities (e.g., information grounding or robust constraint use) rather than inducing indiscriminate instability, consistent with the separation between fine-grained deltas and the overall recovery rate in the figure.

High recovery does not imply uniformly preserved diagnostics. Several models sustain relatively strong recovery while still exhibiting pronounced drops on particular diagnostics, indicating compensatory behavior: outputs can be reconstructed at the sequence level even when intermediate structural signals degrade. Conversely, models with low recovery typically show broader declines across multiple diagnostics, suggesting recovery failures coincide with multi-faceted struc-

Input (Scrambled): 城 (City), 倾 (Topple), 国 (State), 倾 (Topple)

Target: 倾国倾城(Qing Guo Qing Cheng). Meaning: Extreme female beauty.

Qwen-3-8B **Literal Hallucination**

Generated explanation:
指国家和城池被攻破或毁灭，常用于形容重大失败或灾难性后果。*(Refers to a country or city being breached.)*

Error note: Structurally fluent output (high S_{Cons}) but incorrect logic due to literal character-level composition.

DeepSeek-V3 **Precise Retrieval**

Generated explanation:
原指君主过分迷恋女色而亡国。后形容女子容貌极美。*(Originally refers to a monarch losing his kingdom.)*

Error note: Correctly captures the idiomatic metaphorical meaning.

Figure 7: **Case Study on Semantic Hallucination.** Qwen-3-8B produces a fluent but incorrect literal interpretation, while DeepSeek-V3 correctly retrieves the idiomatic metaphor.

tural breakdown rather than a single isolated weakness.

4.7 Case Study: Semantic Hallucination Induced by Literal Compositionality

To illustrate how semantic plausibility can mask reconstruction failure, Figure 7 presents a representative case on *Qing Guo Qing Cheng*. We observe two contrasting behaviors: some models retrieve the idiom-level metaphor correctly, while others interpret characters literally and hallucinate a compositional meaning (e.g., "toppling a country and a city"). This example highlights why OrderProbe requires deterministic reconstruction and diagnostic metrics beyond surface semantic similarity.

4.8 Memorization and Frequency Analysis

A natural concern is whether reconstruction mainly reflects memorization of frequent expressions. We therefore conduct a pilot frequency-stratified analysis (Appendix B.4). Although recovery improves with idiom frequency, reconstruction failures remain common even in high-frequency bins, suggesting that OrderProbe captures structural limitations beyond pure coverage effects.

5 Related Work

Robustness and Adversarial Evaluation The evaluation of model robustness has evolved from surface-level perturbations to semantic integrity. Early benchmarks like TextFlint (Wang et al., 2021) and CheckList (Ribeiro et al., 2020) primarily focused on lexical noise, such as typos and synonym substitution. More recent works have examined robustness in instruction following (Li et al., 2024) and logical reasoning (Lin et al., 2025). Unlike general adversarial attacks that seek to break model performance with noise, OrderProbe probes the cognitive limit of reconstructing canonical order from semantically valid but positionally disordered inputs.

Idiom and Figurative Language Processing Prior research on idioms has predominantly treated them as static lexical units. Datasets such as ChID (Zheng et al., 2019) and Chengyu-Bench (Fu et al., 2025) focus on cloze tests, retrieval, or translation (Donthi et al., 2025). While these benchmarks evaluate whether a model knows an idiom, they do not test whether the model understands its composition. As demonstrated in our experiments, models often exhibit a "knowledge-structure dissociation," performing well on retrieval tasks while failing on the reconstruction tasks proposed in OrderProbe.

Positional Bias and Structural Reasoning Our findings resonate with theoretical studies on the positional limitations of Transformer architectures. Sinha et al. (2021) observed that models are often insensitive to word order in classification tasks, while Berglund et al. (2024) identified the "Reversal Curse," highlighting the inability of models to generalize relationships in reverse directions. OrderProbe provides empirical evidence for these phenomena in the context of East Asian linguistic structures, confirming that even high-resource languages like Chinese are susceptible to severe structural fragility when local semantic anchors are displaced.

6 Conclusion

We introduce **OrderProbe**, a deterministic benchmark for structural reconstruction using four-character expressions in Chinese, Japanese, and Korean. Experiments on twelve LLMs show that exact recovery remains difficult, with zero-shot performance often below 35%, revealing a clear semantic–structure dissociation. We further find that reconstruction success depends strongly on script typology, where logographic scripts provide stronger local anchors and Hangul serves as a strict negative control. We release OrderProbe and our diagnostic framework to support future research on structural robustness in LLMs.

Limitations

Our OrderProbe is designed to evaluate model capabilities through structural reasoning tasks. The types within our dataset reflect our effort to collect data from multiple perspectives. However, we acknowledge that the scope of knowledge covered is inherently limited by the expertise of the contributors and cannot encompass all facets of linguistic structural integrity. Moreover, structural robustness currently lacks a universally accepted definition. Our evaluation is based on six syntactic categories that, while comprehensive, do not cover the full spectrum of linguistic patterns.

Since our dataset is constructed by referencing materials in three languages, we provide multilingual evaluation data. However, the influence of different script typologies on model performance has not been thoroughly analyzed for alphabetic languages. We have not conducted detailed experiments on morphologically rich systems where structural perturbations might yield different results.

In addition, some data points involve crystallized expressions that are present in pre training corpora. We recognize that correct reconstruction depends on the memorization of these fixed patterns. Although the results are accurate within this specific domain, future developments may require models to handle novel combinations beyond memorized content. For these items, we ensure their validity within the current linguistic context. Although these limitations are relevant, we believe they extend beyond the scope and original intent of our work. We hope that future studies will address these challenges in more detail.

Ethics Statement

The OrderProbe dataset integrates linguistic resources from the public domain under fair use principles for academic research. We confirm that the collected expressions are free from copyright restrictions and strictly adhere to data usage policies. All participants provided informed consent and followed a rigorous protocol to exclude offensive or discriminatory content.

Our data construction pipeline utilizes LLMs for semantic augmentation in full compliance with provider terms of service. We declare that all generated outputs underwent human verification to ensure factual accuracy and safety. This work serves as a diagnostic benchmark for enhancing model robustness. We foresee no direct negative societal impacts or misuse risks associated with the release of this dataset.

References

Anthropic. 2024. Claude model report. `https://www.anthropic.com/transparency/model-report`. Accessed: 2025-01.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: LLMs trained on ”a is b” fail to learn ”b is a”. In *I Can’t Believe It’s Not Better Workshop: Failure Modes in the Age of Foundation Models*.

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024. Wenmind: A comprehensive benchmark for evaluating large language models in chinese classical literature and language arts. In *Advances in Neural Information Processing Systems*, volume 37, pages 51358–51410. Curran Associates, Inc.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. Chengyu-bench: Benchmarking large language models for chinese idiom understanding and use. *Preprint*, arXiv:2506.18105.

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

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.

Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. 2024. Evaluating the instruction-following robustness of large language models to prompt injection.

In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 557–568, Miami, Florida, USA. Association for Computational Linguistics.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Tongguan Wang, Mingmin Wu, Guixin Su, Dongyu Su, Yuxue Hu, Zhongqiang Huang, and Ying Sha. 2025. Mchirc: A multimodal benchmark for chinese idiom reading comprehension. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25398–25406. AAAI Press.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, and 15 others. 2021. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Rotbench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. *Preprint*, arXiv:2401.08326.

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.

A Formal Definitions and Rationale of Evaluation Metrics

This appendix details the mathematical formulation and design rationale for the evaluation metrics employed in OrderProbe. To avoid the ambiguity often introduced by aggregated leaderboards, we adopt a **multi-dimensional diagnostic framework**. Each metric is designed to isolate a specific cognitive failure mode—ranging from structural rigidity to semantic hallucination—thereby providing a granular dissection of model capabilities. These metrics are reported independently in the main paper and are not aggregated into a single score.

A.1 Primary Metric: Reconstruction Accuracy

Rationale The ability to recover the canonical form serves as the direct proxy for *Structural Planning*. Unlike semantic comprehension, which can be inferred probabilistically from keywords, structural reconstruction requires resolving specific positional dependencies. Strict matching is necessary to distinguish true structural understanding from partial semantic recall.

Formulation Let x denote a canonical four-character idiom and $x' = \mathcal{T}(x, \pi)$ its internally reordered variant. Given a model output o, reconstruction is marked as correct if o explicitly contains the canonical idiom x or an accepted surface variant.

The Reconstruction Accuracy is defined as:

$$\text{RecAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{x}_i = x_i), \tag{7}$$

In this context, $\mathbb{I}(\cdot)$ represents the indicator function.

A.2 Semantic Accuracy

Rationale and Component Selection Assessing idiom interpretation requires capturing both deep metaphorical meaning and precise entity details. Single-dimensional metrics cannot capture this duality. Vector-based similarity measures often miss specific entity errors, while n-gram overlap fails to recognize valid metaphorical paraphrases. To address this, we construct a tiered hybrid metric, S_{Acc}, that integrates signals from both interaction-based and lexical levels.

Weighting Justification We assign weights to reflect the functional hierarchy of each verification method. We treat the cross-encoder as the primary evaluator and assign it a weight of 0.5. Unlike bi-encoder architectures, which compress sentences into independent vectors, cross-encoders perform full self-attention over the concatenated input. This enables them to detect subtle semantic entailments that are essential for precise interpretation. We assign a moderate weight of 0.3 to an ensemble of embedding-based measures, including BERTScore, STS, and cosine similarity, which serves as a robust semantic baseline. Relying on a single embedding model may introduce architectural bias. Integrating distinct representation approaches reduces variance and provides a more general measure of semantic proximity. We allocate the remaining weight of 0.2 to the F_β score as a lexical safeguard. Models can exhibit high similarity scores while hallucinating or mismatching specific entities. This lexical component therefore penalizes factual discrepancies.

Formulation For an explanation e and a reference definition r, the composite Semantic Accuracy score is defined as:

$$\begin{aligned} S_{\text{Acc}} = w_1 \cdot S'_{\text{ce}} + w_2 \cdot \left(\frac{S'_{\text{bert}} + S'_{\text{sts}} + S'_{\text{cos}}}{3} \right) \\ + w_3 \cdot S_{f\beta}. \end{aligned} \tag{8}$$

S'_{ce} denotes the normalized score produced by the cross-encoder. The term in parentheses is the averaged embedding similarity score. $S_{f\beta}$ denotes the lexical F_β score. We set the weights to $w_1 = 0.5$, $w_2 = 0.3$, and $w_3 = 0.2$ based on the hierarchy described above.

A.3 Structural Consistency

Rationale and Design Philosophy Characterizing structural stability requires disentangling a model's latent recognition capability from its execution volatility under noise. For any given permuted input, the model may generate multiple plausible explanations. We assume that the peak semantic score $S_{\max}$ among these generations reflects an upper bound on the model's recognition capability. In contrast, the mean semantic score S_{mean} reflects the expected realized performance. The gap between this potential maximum and the expected mean quantifies the *internal volatility* for that specific structural configuration. To capture overall stability, we aggregate these gaps across the entire permutation space. We adopt a multiplicative aggregation strategy to compute the final consistency score. This dual-penalty design ensures that a robust model must minimize both the average capability loss and the sensitivity of that loss to structural changes. The multiplicative formulation further ensures that a deficiency in either dimension disproportionately penalizes the final consistency metric.

Formulation We operationalize structural consistency through three interlinked equations. Let $\mathcal{P}$ denote the complete set of permutations applied to a specific idiom. For each permutation $p \in \mathcal{P}$, the model generates a set of explanations. Let S_{mean} be the average semantic accuracy score over these explanations, and let $S_{\max}$ be the maximum semantic accuracy score among them. We define the *performance deviation* E_{perf} as the average gap between recognition capability and realized performance, aggregated across all permutations:

$$E_{\text{perf}} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} (S_{\max} - S_{\text{mean}}). \tag{9}$$

We define the *rigidity sensitivity* R_{sens} as the maximum gap observed across the permutation space:

$$R_{\text{sens}} = \max_{p\in\mathcal{P}} (S_{\max} - S_{\text{mean}}). \tag{10}$$

Finally, we compute the composite consistency score S_{Cons} as the product of the retention rates derived from both deviations:

$$S_{\text{Cons}} = (1 - E_{\text{perf}}) \times (1 - R_{\text{sens}}). \tag{11}$$

This formulation enforces structural stability by penalizing internal volatility in a mathematically explicit manner.

A.4 Degradation Under Internal Reordering

Rationale and Normalization Logic Isolating the impact of structural noise requires a performance baseline for each specific idiom. We use the score obtained from the canonical non-reordered input, S_{orig}, as the reference capability. A low score on a scrambled input indicates structural fragility only if the model demonstrates competence on the canonical form. Direct comparison without normalization fails to distinguish between structural sensitivity and general knowledge deficits. Consequently, we quantify degradation as a relative loss ratio. MDR captures the expected utility loss under random structural perturbations. MDA captures a lower bound on performance reliability, reflecting the worst-case structural collapse. The maximization term in MDA extracts the worst-case signal across the permutation group to maintain a consistent definition.

Formulation We formalize the degradation metrics through the following equations. $\mathcal{P}$ denotes the set of permutations applied to the idiom. S_{orig} represents the semantic accuracy score of the model on the canonical input. S_{pert}^{p} denotes the score under permutation p. We define the Mean Degradation Rate (MDR) as the average relative performance drop across all permutations.

$$\text{MDR} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{S_{\text{orig}} - S_{\text{pert}}^{p}}{S_{\text{orig}}} \tag{12}$$

We define the Maximum Degradation Amount (MDA) as the maximum relative drop observed across the permutation group.

$$\text{MDA} = \max_{p \in \mathcal{P}} \left(\frac{S_{\text{orig}} - S_{\text{pert}}^{p}}{S_{\text{orig}}} \right) \tag{13}$$

These values quantify the average and worst-case loss of capability attributable to internal reordering.

A.5 Robustness Metrics

Rationale and Dual Dimensionality We assess robustness along two perturbation dimensions: Sequential Robustness and Structural Robustness. Sequential Robustness measures the resilience of the model against internal ordering permutations. Structural Robustness evaluates the stability of performance across distinct syntactic categories. We select the mean semantic accuracy $S_{\text{Acc}}^{\text{mean}}$ as the basis for these metrics. While maximum scores reflect peak recovery capability, average scores capture stability under stochastic degradation and thus provide a stricter and more sensitive indicator of robustness. The final metric integrates these dimensions using a harmonic mean, ensuring that high robustness requires resilience to both combinatorial scrambling and syntactic variation.

Sequential Robustness Formulation We quantify Sequential Robustness S_{seq} by combining the mean and maximum degradation rates defined in Appendix A.4. We define S_{seq} as the complement of a weighted sum of these degradations. We assign equal weights $\alpha = 0.5$ and $\beta = 0.5$ to balance average stability and limit-state resilience.

$$S_{\text{seq}} = 1 - (\alpha \cdot \text{MDR} + \beta \cdot \text{MDA}) \tag{14}$$

Structural Robustness Formulation We quantify Structural Robustness S_{struct} by measuring performance variance across syntactic categories. Let T denote the set of six syntactic structure types indexed by k. For the k-th structure, we compute the intra-group mean accuracy μ_k over the idiom set D_k.

$$\mu_k = \frac{1}{|D_k|} \sum_{x \in D_k} S_{\text{Acc}}^{\text{mean}}(x) \tag{15}$$

We calculate the standard deviation σ of these group means to capture the sensitivity of the model to syntactic form. We define S_{struct} as the complement of the normalized deviation, ensuring that consistent performance across diverse structures yields a high score.

$$S_{\text{struct}} = 1 - \text{Normalize}(\sigma(\mu_1, \mu_2, ..., \mu_6)) \tag{16}$$

Composite Robustness Score We compute the final Robustness Score S_{Rob} as the harmonic mean of the sequential and structural components. This aggregation penalizes imbalance, requiring the model to demonstrate robustness in both internal ordering and syntactic generalization.

$$S_{\text{Rob}} = \frac{2 \cdot S_{\text{seq}} \cdot S_{\text{struct}}}{S_{\text{seq}} + S_{\text{struct}}} \tag{17}$$

A.6 Logical Validity

Rationale and Entailment Check High lexical similarity does not guarantee factual correctness.

Models may generate explanations that share keywords with the target while conveying contradictory meanings through negation or causal inversion. Lexical metrics fail to detect these logical errors. To address this, we employ a Natural Language Inference (NLI) model as a validity filter. This component evaluates whether the generated explanation logically entails the canonical definition, ensuring that high scores reflect true semantic understanding rather than superficial keyword matching.

Formulation We quantify logical validity S_{Log} as the probability of entailment predicted by a pretrained NLI model. Given a generated explanation e and a reference definition r, we compute the entailment probability:

$$S_{\text{Log}} = P_{\text{NLI}}(e \Rightarrow r) \tag{18}$$

A.7 Information Density

Rationale and Verbosity Penalty Models frequently employ knowledge-dumping strategies, generating excessively verbose explanations to maximize keyword coverage. This behavior inflates recall-based metrics without reflecting precise understanding. To counteract this, we introduce an Information Density metric that rewards high information content per token. We impose a brevity penalty (BP) on outputs that exceed the reference length without proportional semantic gain. This mechanism encourages the model to prioritize concise and accurate reasoning over redundant generation.

Formulation We compute Information Density S_{Info} by scaling the precision score with a length-based penalty:

$$S_{\text{Info}} = \text{BP} \times P_{\text{ROUGE}} \tag{19}$$

P_{ROUGE} denotes the ROUGE precision score. BP represents the brevity penalty factor, downweighting explanations that exhibit low information density relative to the canonical definition.

A.7.1 Human Audit Rubrics

We conduct a targeted human audit to validate the proposed multi-dimensional evaluation framework. The audit assesses whether the automated dimensions align with expert judgments under explicit and controlled rating criteria.

For each language, we sample a fixed subset of instances from the evaluation set without conditioning on model performance or automated metric values. Each instance is independently annotated by five annotators who are proficient in the corresponding language. Annotators evaluate model-generated explanations without access to automated scores.

Human evaluation follows predefined rubrics that map qualitative judgments to normalized scores. All dimensions are rated independently.

Accuracy score s^{mean}_{Acc}. A score of at least 0.8 indicates a fully correct explanation. A score between 0.6 and 0.8 indicates that the main meaning is correct with minor inaccuracies. A score between 0.4 and 0.6 indicates noticeable errors despite a largely correct meaning. A score below 0.4 indicates severe errors and an incorrect explanation.

Logical validity score s_{Logic}. This score measures whether the explanation is logically consistent with the canonical definition (i.e., it entails the reference meaning and does not introduce contradictions). A score of at least 0.8 indicates that the explanation matches the reference meaning without logical conflicts. A score between 0.6 and 0.8 indicates that the main meaning is correct but includes minor logical omissions or slight mis-specifications. A score between 0.4 and 0.6 indicates mixed correctness, with partial contradictions or incorrect causal relations. A score below 0.4 indicates a logically inconsistent or contradictory explanation.

Aggregation and agreement. Final human audit scores are obtained by averaging annotator ratings. Inter-annotator agreement is moderate to high across languages.

A.8 Human Audit Protocol

We conduct a targeted human audit to validate the proposed multi-dimensional evaluation framework. The audit assesses whether automated evaluation dimensions align with expert judgments under explicit and controlled rating criteria.

For each language, we sample a fixed subset of instances from the evaluation set without conditioning on model performance or automated metric values. The audit includes 290 instances for ZH-CN, 430 instances for ZH-TW, 170 instances for JA, and 110 instances for KO. Each instance is independently annotated by five annotators who are proficient in the corresponding language. Annotators evaluate model-generated explanations without access to automated scores. Final human scores are

Language	N	ρ_{Acc}^{mean}	ρ_{Logic}	ρ_{Info}
ZH-CN	290	0.787	0.822	0.743
ZH-TW	430	0.754	0.801	0.717
JA	170	0.694	0.739	0.665
KO	110	0.723	0.762	0.685

Table 5: Instance level Spearman rank correlation between human audit ratings and automated metric scores on the audited samples.

obtained by averaging annotator ratings, and inter-annotator agreement is moderate to high across languages.

Human evaluation follows predefined rubrics that map qualitative judgments to normalized scores. All dimensions are rated independently.

Accuracy score s_{Acc}^{mean}. A score of at least 0.8 indicates a fully correct explanation. Scores between 0.6 and 0.8 indicate that the main meaning is correct with minor inaccuracies. Scores between 0.4 and 0.6 indicate noticeable errors despite a largely correct meaning. Scores below 0.4 indicate severe errors and an incorrect explanation.

Logical quality score s_{Logic}. A score of at least 0.8 indicates a correct and concise explanation with no redundant content. Scores between 0.6 and 0.8 indicate mostly correct explanations with some redundancy. Scores between 0.4 and 0.6 indicate substantial redundancy with mixed correctness. Scores below 0.4 indicate incorrect explanations or predominantly redundant content.

Information score s_{Info}. The information score reflects the overall information density and usefulness of the explanation. Higher scores indicate more informative and content-rich explanations, while lower scores indicate limited or insufficient information.

To quantify alignment at the instance level, we compute Spearman's rank correlation between human audit ratings and automated metric scores on the same audited samples. Correlations are reported for all dimensions that produce per-instance scores under both human and automated evaluation. Table 5 summarizes the alignment results across languages.

A.9 Alignment Between Human Evaluation and Automated Metrics

This section examines the alignment between human expert judgments and the proposed automated multi-dimensional evaluation framework. Table 6 compares aggregated human audit scores with automated metric outputs across languages.

Across all evaluated languages, the automated metrics exhibit consistent directional alignment with human assessments. Dimensions related to logical coherence, robustness, and internal consistency show similar relative patterns under both evaluation methods, suggesting that the metrics capture structured properties of model behavior rather than superficial numerical differences.

Differences in absolute scores are observed between human audits and automated evaluations. Such variation is expected given the interpretive nature of human judgment and the formal constraints of automated scoring. Importantly, the relative ordering of evaluation dimensions remains stable across methods, indicating that the proposed multi-dimensional framework preserves the structure of human evaluation across languages.

Language	Evaluation Method	s_{Acc}^{mean}	s_{Cons}	s_{Logic}	s_{Rob}	s_{Info}
ZH-CN	Human Audit	0.412	0.758	0.829	0.715	0.456
	Automated Metric	0.482	0.663	0.749	0.625	0.516
ZH-TW	Human Audit	0.405	0.731	0.779	0.701	0.336
	Automated Metric	0.480	0.641	0.699	0.613	0.406
JA	Human Audit	0.413	0.654	0.632	0.708	0.547
	Automated Metric	0.479	0.559	0.552	0.621	0.617
KO	Human Audit	0.353	0.875	0.443	0.698	0.329
	Automated Metric	0.413	0.788	0.363	0.604	0.419

Table 6: Comparison between human expert audits and automated multi dimensional evaluation scores across languages.

A.10 Generation Protocol and Nondeterminism

All models are evaluated using inference configurations consistent with those commonly adopted in large-scale benchmarking studies. For models that expose a temperature parameter, we set the temperature to 0 to enforce deterministic decoding. No task-specific tuning or customized decoding strategies are applied.

Open-source models are run locally using official checkpoints and standard inference pipelines. Closed-source models are accessed through their official APIs, following the default configurations recommended by their respective providers. All API-based evaluations are conducted between October 2025 and January 2026. When configurable parameters are available, we align them across models to the extent possible; otherwise, we retain provider defaults.

Across all experiments, we use identical prompt templates and input perturbations. No fine-tuning, reranking, or post-processing is performed. This deployment protocol ensures that performance differences primarily reflect intrinsic model behaviors rather than artifacts of inference configuration, stochastic decoding, or serving infrastructure.

B Appendix: Details of Semantic and Logical Evaluation

This appendix specifies implementation details of the automated evaluation metrics for semantic explanation quality and logical consistency to ensure reproducibility and cross-lingual comparability.

B.1 Semantic Explanation Scoring

Semantic explanation quality is measured using a composite score that integrates interaction-based semantic matching, representation-based similarity, and content coverage.

For interaction-based matching, we use the cross-encoder reranker BAAI/bge-reranker-base. Given a model-generated explanation and a set of reference explanations, the cross-encoder produces a relevance score for each explanation–reference pair. These scores are normalized to a bounded scale and aggregated across multiple references using either max or mean aggregation. The same cross-encoder model is applied uniformly across all languages.

For representation-based similarity, we use sentence embedding models to compute cosine similarity between explanations and references. For Simplified Chinese and Traditional Chinese, we use BAAI/bge-small-zh-v1.5. For Japanese and Korean, we use sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. These models are selected for their multilingual coverage and efficiency. No translation is applied, and all texts are evaluated in their original language.

In addition, we optionally include BERTScore F1 as a token-level semantic similarity signal, using language-specific settings for Chinese, Japanese, and Korean. Lexical overlap features are treated as auxiliary signals and do not dominate the final score.

The final semantic accuracy score is computed as a weighted linear combination of cross-encoder similarity, representation-based similarity, and the content-word F_β score, with fixed weights shared across languages.

B.2 Logical Consistency via Natural Language Inference

Logical consistency between model explanations and reference explanations is measured using the multilingual natural language inference (NLI) classifier MoritzLaurer/mDeBERTa-v3-base-mnli-xnli. The same NLI model is applied to all languages without translation.

For each reference explanation, we compute NLI in two directions: from the reference to the model output and from the model output to the reference. This bidirectional evaluation reduces asymmetry caused by verbosity or paraphrasing. We use the entailment probability as the primary signal, while the contradiction probability serves as a secondary penalty when entailment evidence is present.

Specifically, for each reference, the score is defined as the maximum entailment probability across the two directions, with an additional contradiction-aware adjustment activated only when entailment exceeds a small threshold. Scores are aggregated across multiple references using max aggregation. This reflects the assumption that a model explanation is logically consistent if it entails at least one valid reference explanation.

The NLI scores are treated as continuous consistency signals and are not converted into binary decisions. No post-hoc calibration is applied, and scores are used only for relative comparison and averaging.

B.3 Reference Construction and Evaluation Scope

Reference explanations consist of canonical dictionary definitions augmented with paraphrased variants to cover legitimate surface variation. Dictionary definitions serve as the primary semantic anchors, while paraphrased references reduce sensitivity to wording differences. All evaluation metrics aggregate over multiple references to avoid dependence on any single phrasing.

We evaluate the main reconstruction accuracy metric alongside semantic and logical metrics. While some items may appear in pretraining corpora, reconstruction accuracy is not interpreted in isolation. Instead, differences across semantic, logical, and robustness metrics are used to assess structure-aware semantic grounding beyond surface recall.

Frequency Bin	Recovery (%)
High	**30.576**
Mid	**19.794**
Low	**10.231**

Table 7: **Pilot Frequency-Stratified Recovery (Zero-shot).** Exact recovery rates (%) across high-, mid-, and low-frequency bins.

B.4 Frequency-Stratified Recovery Pilot

A potential concern is that exact recovery on four-character expressions may be influenced by memorization, since frequent idioms are more likely to appear in pretraining corpora. To quantify this effect, we conduct a pilot frequency-stratified analysis under the zero-shot setting using two representative models that span a wide capability range: GPT-4o and Qwen3-8B. We stratify expressions into three frequency bins (high, mid, and low) and measure the recovery rate as exact canonical reconstruction. Table 7 reports bin-wise recovery averaged across the two models.

As shown in Table 7, we observe a clear monotonic frequency effect: recovery is highest on high-frequency expressions, decreases on mid-frequency expressions, and remains non-trivial on low-frequency expressions. This indicates that lexical frequency contributes to reconstruction performance, while meaningful recovery persists even in the low-frequency regime.

C Prompt Templates for Idiom Meaning Inference

All baseline runs use a fixed two-message prompting scheme consisting of a system instruction and a user query template. The system instruction constrains the output format to a single-line response in Chinese. The user template provides the target idiom and repeats the required output schema to reduce formatting deviations. Across different experimental conditions, only the prompt text is modified, while the model, decoding settings, batch processing logic, and output parsing rules remain unchanged. For multilingual settings, the same prompting structure is retained, with the prompt content adapted through direct translation or equivalent reformulation while preserving identical functional roles and formatting requirements.

To facilitate reproducibility, Table 8 presents the baseline prompt templates used for idiom meaning inference across languages. The prompts are shown in their actual formatted layout, including line breaks, to illustrate the structural constraints imposed on model outputs. No task-specific examples, demonstrations, or additional guidance are included beyond the required format specification.

All experiments are conducted using fixed inference settings. For the *baseline* condition, no task-specific examples or chain-of-thought instructions are provided beyond the one-line format constraint. For the *CoT* condition, we use a fixed two-line CoT template (Table 10) while keeping all decoding settings identical.

D Semantic Hallucination vs. Idiomatic Retrieval

Figure 7 presents a representative failure case on the idiom *Qing Guo Qing Cheng*. Although the scrambled characters preserve strong lexical cues, some models still produce fluent but incorrect literal interpretations, reflecting compositional hallucination. In contrast, stronger models correctly retrieve the idiomatic metaphor, demonstrating that semantic plausibility and structural reconstruction can diverge sharply.

E Qualitative Analysis and Case Studies

In this section, we provide a qualitative visual analysis of model performance across different experimental settings. Figures 8 and 9 illustrate baseline (zero-shot) behaviors and highlight common failure modes such as hallucination and structural blindness. Figures 10 through 13 further compare Chain-of-Thought (CoT) and few-shot strategies for each language (ZH-CN, ZH-TW, JP, KO). These visualizations display the system prompts, user queries, and model outputs, and they illustrate the strict constraints imposed by our evaluation pipeline.

F Reference Performance on Canonical Inputs

To isolate the specific impact of structural perturbation, we conducted a control experiment using the canonical (unperturbed) forms of the expressions. This serves two critical purposes: establishing a performance upper bound for each model and validating the reliability of our automated diagnostic metrics.

Table 11 reports the diagnostic scores under this ideal condition. We observe that all evaluated models demonstrate high proficiency when the input structure is intact. For instance, semantic fidelity scores (S_{Acc}^{mean}) consistently range between 0.75

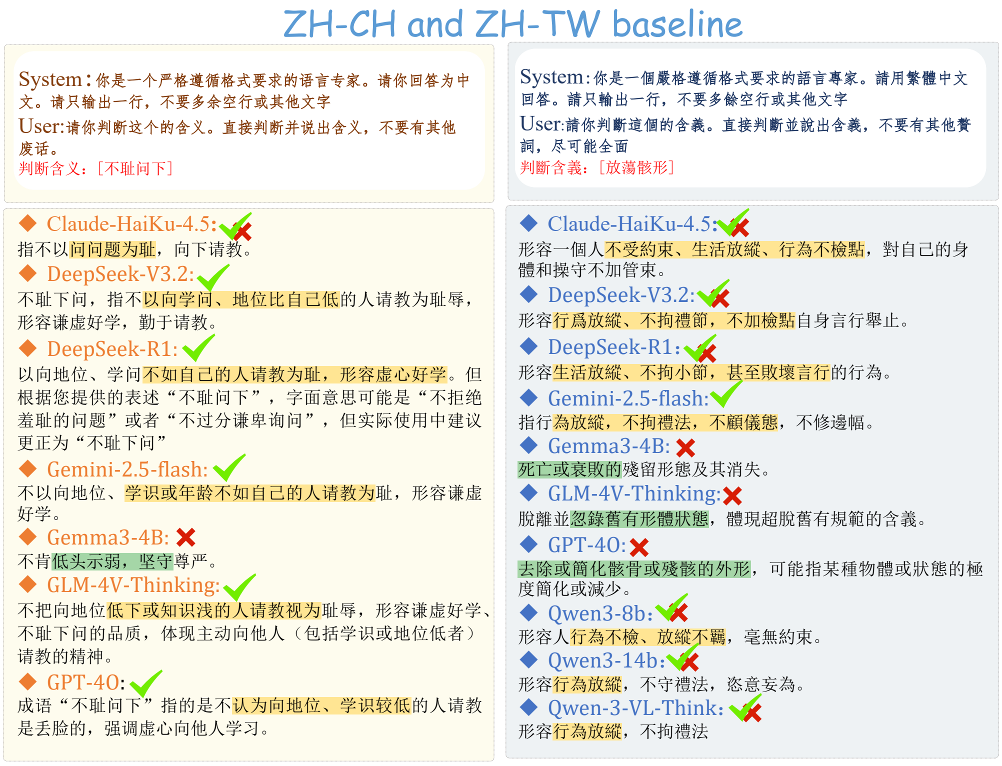
and 0.88 across different model tiers, and logical validity (S_{Logic}) remains high.

These results support two key conclusions regarding the main experiments:

1. **Knowledge Verification:** The high baseline performance confirms that the models possess the requisite lexical knowledge in their pretraining distribution. Therefore, the failures observed in the scrambled settings (Section 4.2) cannot be attributed to a lack of semantic understanding, but rather to a specific deficit in structural reconstruction.

2. **Metric Validation:** The high scores under canonical conditions validate our evaluation pipeline. They indicate that our automated metrics accurately reward correct semantic interpretations, ruling out the possibility that the low scores in the main experiments are artifacts of an overly punitive scoring mechanism.

The stark contrast between this high-performance baseline and the significant degradation observed under permutation (where recovery rates frequently drop below 35%) provides compelling evidence for the semantic–structure dissociation hypothesis discussed in the main text.



ZH-CH and ZH-TW baseline

System：你是一个严格遵循格式要求的语言专家。请你回答为中文。请只输出一行，不要多余空行或其他文字
User:请你判断这个的含义。直接判断并说出含义，不要有其他废话。
判断含义：［不耻问下］

- Claude-HaiKu-4.5: ✓✗
指不以问问题为耻，向下请教。
- DeepSeek-V3.2: ✓
不耻下问，指不以向学问、地位比自己低的人请教为耻辱，形容谦虚好学，勤于请教。
- DeepSeek-R1: ✓
以向地位、学问不如自己的人请教为耻，形容虚心好学。但根据您提供的表述“不耻问下”，字面意思可能是“不拒绝羞耻的问题”或者“不过分谦卑询问”，但实际使用中建议更正为“不耻下问”
- Gemini-2.5-flash: ✓
不以向地位、学识或年龄不如自己的人请教为耻，形容谦虚好学。
- Gemma3-4B: ✗
不肯低头示弱，坚守尊严。
- GLM-4V-Thinking: ✓
不把向地位低下或知识浅的人请教视为耻辱，形容谦虚好学、不耻下问的品质，体现主动向他人（包括学识或地位低者）请教的精神。
- GPT-4O: ✓
成语“不耻问下”指的是不认为向地位、学识较低的人请教是丢脸的，强调虚心向他人学习。

System:你是一個嚴格遵循格式要求的語言專家。請用繁體中文回答。請只輸出一行，不要多餘空行或其他文字
User:請你判斷這個的含義。直接判斷並說出含義，不要有其他贅詞，尽可能全面
判斷含義：［放蕩骸形］

- Claude-HaiKu-4.5: ✓✗
形容一個人不受約束、生活放縱、行為不檢點，對自己的身體和操守不加管束。
- DeepSeek-V3.2: ✓✗
形容行爲放縱、不拘禮節，不加檢點自身言行舉止。
- DeepSeek-R1: ✓✗
形容生活放縱、不拘小節，甚至敗壞言行的行為。
- Gemini-2.5-flash: ✓
指行為放縱，不拘禮法，不顧儀態，不修邊幅。
- Gemma3-4B: ✗
死亡或衰敗的殘留形態及其消失。
- GLM-4V-Thinking: ✗
脫離並忽錄舊有形體狀態，體現超脫舊有規範的含義。
- GPT-4O: ✗
去除或簡化骸骨或殘骸的外形，可能指某種物體或狀態的極度簡化或減少。
- Qwen3-8b: ✓✗
形容人行為不檢、放縱不羈，毫無約束。
- Qwen3-14b: ✓✗
形容行為放縱，不守禮法，恣意妄為。
- Qwen-3-VL-Think: ✓✗
形容行為放縱，不拘禮法

Figure 8: **Baseline performance on Chinese scripts illustrating Simplified Chinese on the left and Traditional Chinese on the right.** Even powerful models often fail to recover the canonical order.

JP and KO baseline

System：너는 형식을 엄격히 지키는 언어 전문가야. 한국어로만 답해,반드시 한 줄로만 출력해. 다른 말 금지.
User:내가 보내는 내용의 의미를 한 줄로만 말해. 군말 금지, 가능한 한 완전하게 설명해.
내용：［개관정사］

- Claude-HaiKu-4.5: ✓✗
전체적인 상황이나 현황을 살펴보고 판단하는 것.
- DeepSeek-V3.2: ✗
특정 목적을 위한 인력 자원 상태를 평가하고 확인하는 과정.
- DeepSeek-R1: ✗
속성 명세와 상황 판단을 합친 의미다.
- Gemini-2.5-flash:
어떤 대상의 숨겨진 본질이나 핵심을 철저히 파헤쳐 상세히 알아내거나 정밀하게 평가하는 것을 비유적으로 이르는 말.
- Gemma2-7B: ✗
개관사정.
- GLM-4V-Think: ✗
개관사정은 개시 과정과 관련된 상황이나 사정, 또는 이유를 의미합니다.
- GPT-4O: ✓✗
전체적인 상황이나 내용을 대략적으로 살펴보고 판단하는 것.
- Qwen3-8b: ✓✗
개관사정은 특정 시점의 현황이나 상황을 전달하는 말로, 현재의 상태나 상황을 간략히 설명하는 데 사용된다.

System：あなたは形式を厳守する言語専門家です。日本語で答えてください,必ず1行だけ出力し、余計な文言・空行……
User:これから送る内容の意味を一行だけで答えてください。余計な文言は禁止。できるだけ完全に説明してください。
内容：［画竜睛点］

- Claude-HaiKu-4.5: ✗
田にん字は重要な点を指し示すことを巧みに表現する。
- DeepSeek-V3.2: ✗
絵の仕上げにとどめを刺すこと。
- DeepSeek-R1: ✗
画竜睛点は冗談で場を和ませるおなかも軽いからかいです。
- Gemini-2.5-flash: ✓
物事の最も肝心な部分を補い、全体を完璧に仕上げて生命を吹き込むこと。
- Gemma2-7B: ✗
龍の目を一点に集めること、すなわち龍を凝視すること。
- GLM-4V-Think: ✓
作品を完成させるためにもっとも重要な要素を最終的に加える行為、全体を一層に見事にする。
- GPT-4O: ✓
画竜睛点（がりょうせいてん）は、絵画や文章で、重要な部分に手を加えて完成度を高めることを指す表現。
- Qwen3-8b: ✓
画竜睛点とは、龍を描くときに最後に目を描くことにより、作品に生き気機が生まれることを指し、物事を完璧にする最後の重要な要素を意味する。

Figure 9: **Baseline performance on Japanese and Korean scripts showing Korean Hangul on the left and Japanese Kanji on the right.** The structural scrambling effectively disrupts local semantic anchors.

ZH-CH— COT and Few-Shot

System：你是一个严谨的逻辑分析专家。请严格遵循用户的指令格式输出结果。
User:请分析以下文本的深层含义。这是一个测试语言逻辑重组能力的认为……
判断含义：［不耻问下］

◆ Claude-HaiKu-4.5(COT): ✓✗
指不以向下问问题为耻，向下请教。
◆ DeepSeek-V3.2(COT): ✓
教导人们要谦虚好学，不以向地位、学识或年龄低于自己的人请教为耻。
◆ DeepSeek-R1(COT): ✓✗
不以向地位低下者请教为耻，体现谦虚好学的精神。
◆ Gemini-2.5-flash(COT):
不应以向地位、学识或经验不如自己的人请教问题为耻，体现了虚心求教、不耻下问的谦逊学习精神，鼓励人们放下身份偏见，勇于从任何人身上汲取知识。✓
◆ Gemma3-4B(COT): ✗
表达了即使因提问而感到羞耻，仍然坚持不懈地寻求真理或帮助的决心与挣扎。
◆ GLM-4V-Thinking(COT): ✓
不以向地位低下者或者不如自己者询问为羞耻，彰显谦虚好学之精神。
◆ GPT-4O(COT): ✓✗
不以向低级别的人请教为耻。

System：你是一个严谨的逻辑分析专家。请严格遵循用户的指令格式输出结果。
User:请仔细阅读以下文本分析示例，学习其分析逻辑和输出格式。然后，对最后给出的待分析文本进行同样的分析……
判断含义：［不耻问下］

◆ Claude-HaiKu-4.5(Few-Shot): ✓✗
比喻谦虚求教，不因为对方地位低而自觉丢脸，强调虚心好学的品德。
◆ DeepSeek-V3.2(Few-Shot): ✓
比喻谦虚好学，不以向地位、学识较低的人请教为耻。
◆ DeepSeek-R1(Few-Shot): ✓✗
形容虚心学习，不以向他人请教为羞耻。
◆ Gemini-2.5-flash(Few-Shot): ✓
指不以向地位或学问比自己低的人请教为耻。形容谦虚好学。
◆ Gemma3-4B(Few-Shot): ✗
比喻骄傲自大，不愿向别人求助。
◆ GLM-4V-Thinking(Few-Shot): ✓✗
形容谦虚好学，不以向地位低下的人请教问题而感到羞耻。
◆ GPT-4O(Few-Shot): ✓✗
比喻不以向比自己地位低的人请教为耻，反而是值得提倡的行为。
◆ Qwen3-8b(Few-Shot): ✓✗
比喻不以向地位低的人请教为耻，形容谦虚好学。

Figure 10: **Comparison of prompting strategies on Simplified Chinese.** Chain of Thought prompting elicits better reasoning but still suffers from hallucination during complex reordering.

ZH-TW— COT and Few-Shot

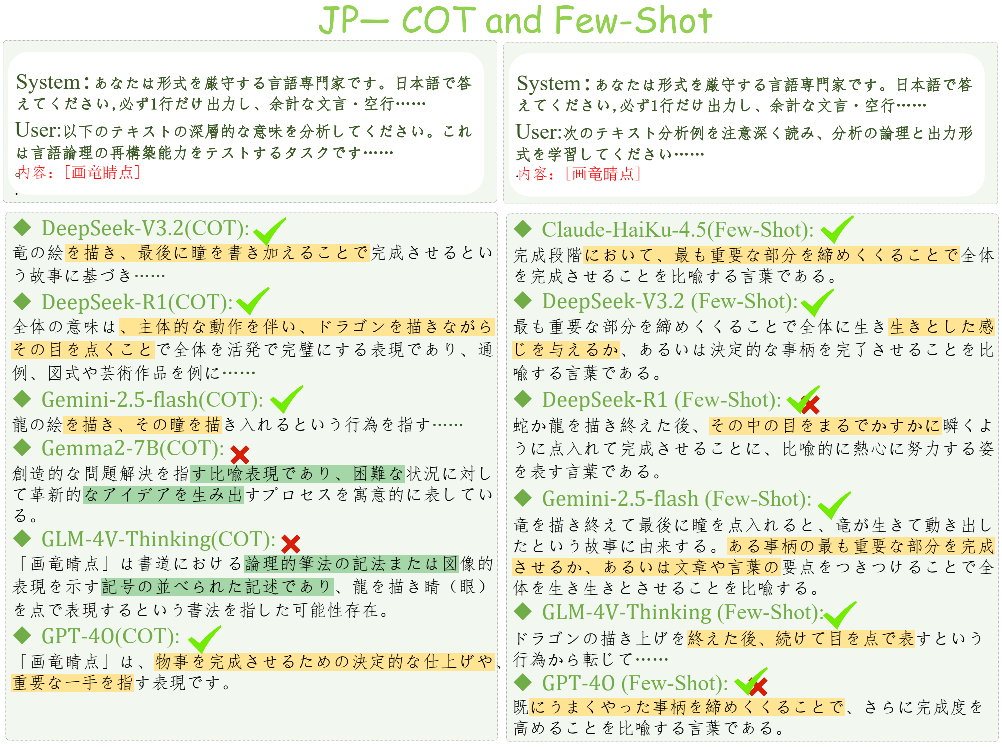
System:你是一個嚴謹的邏輯分析專家。請嚴格遵循用戶的指令格式輸出結果
User:請分析以下文本的深層含義。這是一個測試語言邏輯重組能力的任務……
判斷含義：［脫略骸形］

◆ Claude-HaiKu-4.5(COT): ✓
形容人灑脫不羈，不被世俗的禮法、外在的形體規範所拘束，行事隨性自然，不矯揉造作。
◆ DeepSeek-V3.2(COT) : ✓✗
指行為放縱不羈、不拘泥於世俗禮法或外在形體束縛，追求精神上的超脫與自由。
◆ DeepSeek-R1(COT): ✓
深層含義是象徵性地表示通過忽略或掙脫外在身體和形式的束縛，達成精神上的淨化與自我超越。
◆ Gemini-2.5-flash(COT): ✓
意指超脫、擺脫肉體軀殼的束縛，不為形體、外貌或物質所拘泥，達致精神上的自由、灑脫與豁達。
◆ Gemma3-4B(COT): ✗
指昔日好風度，現在衰老後失去風度。
◆ GLM-4V-Thinking(COT): ✓✗
形容人放縱無度、形骸放蕩的行為與狀態。
◆ GPT-4O(COT): ✓✗
形容人的行為不檢、放縱不羈，毫無約束。
◆ Qwen3-8b(COT): ✗
脫略骸形意指去除外在形體與骨架，強調超越物質表象，追求內在本質或精神層面的自由與純粹。

System:你是一個嚴謹的邏輯分析專家。請嚴格遵循用戶的指令格式輸出結果
User:請仔細閱讀以下文本分析範例，學習其分析邏輯和輸出格式。然後，對最後給出的待分析文本進行同樣的分析……
判斷含義：［脫略骸形］

◆ Claude-HaiKu-4.5(Few-Shot): ✓✗
比喻不為物質身體所累，精神超越世俗，達到忘卻肉身束縛的精神境界。
◆ DeepSeek-V3.2(Few-Shot) : ✓✗
形容人的行為放蕩不羈，不拘禮節，也指不為世俗所拘泥。
◆ DeepSeek-R1(Few-Shot): ✓
形容超越物質形體束縛，追求精神解脫與純粹境界。
◆ Gemini-2.5-flash(Few-Shot): ✓
指不拘泥於形體、外表或世俗禮節，行為灑脫不羈。
◆ Gemma3-4B(Few-Shot): ✗
比喻形式不合乎實際，虛構形狀。
◆ GLM-4V-Thinking(Few-Shot): ✓✗
比喻超脫骸骨形狀拘束並略去相關關係，呈現出豁達自由的精神態勢。
◆ GPT-4O(Few-Shot): ✓✗
比喻脫離或省略外在形象，可能指心境或形態上的變化。
◆ Qwen3-8b(Few-Shot): ✗
比喻超脫形骸、忘卻軀體，形容心神專注或境界高遠。
◆ Qwen3-14B(Few-Shot): ✓
形容不受外在形體的束縛，追求精神或內在的自由。

Figure 11: **Comparison of prompting strategies on Traditional Chinese.** The results highlight significant performance gaps across different inference methods.



JP— COT and Few-Shot

System：あなたは形式を厳守する言語専門家です。日本語で答えてください,必ず1行だけ出力し、余計な文言・空行……
User:以下のテキストの深層的な意味を分析してください。これは言語論理の再構築能力をテストするタスクです……
内容：［画竜睛点］

◆ **DeepSeek-V3.2(COT):** ✓
竜の絵を描き、最後に瞳を書き加えることで完成させるという故事に基づき……

◆ **DeepSeek-R1(COT):** ✓
全体の意味は、主体的な動作を伴い、ドラゴンを描きながらその目を点くことで全体を活発で完璧にする表現であり、通例、図式や芸術作品を例に……

◆ **Gemini-2.5-flash(COT):** ✓
龍の絵を描き、その瞳を描き入れるという行為を指す……

◆ **Gemma2-7B(COT):** ✗
創造的な問題解決を指す比喩表現であり、困難な状況に対して革新的なアイデアを生み出すプロセスを寓意的に表している。

◆ **GLM-4V-Thinking(COT):** ✗
「画竜睛点」は書道における論理的筆法の記法または図像的表現を示す記号の並べられた記述であり、龍を描き睛（眼）を点で表現するという書法を指した可能性存在。

◆ **GPT-4O(COT):** ✓
「画竜睛点」は、物事を完成させるための決定的な仕上げや、重要な一手を指す表現です。

System：あなたは形式を厳守する言語専門家です。日本語で答えてください,必ず1行だけ出力し、余計な文言・空行……
User:次のテキスト分析例を注意深く読み、分析の論理と出力形式を学習してください……
内容：［画竜睛点］

◆ **Claude-HaiKu-4.5(Few-Shot):** ✓
完成段階において、最も重要な部分を締めくくることで全体を完成させることを比喩する言葉である。

◆ **DeepSeek-V3.2 (Few-Shot):** ✓
最も重要な部分を締めくくることで全体に生き生きとした感じを与えるか、あるいは決定的な事柄を完了させることを比喩する言葉である。

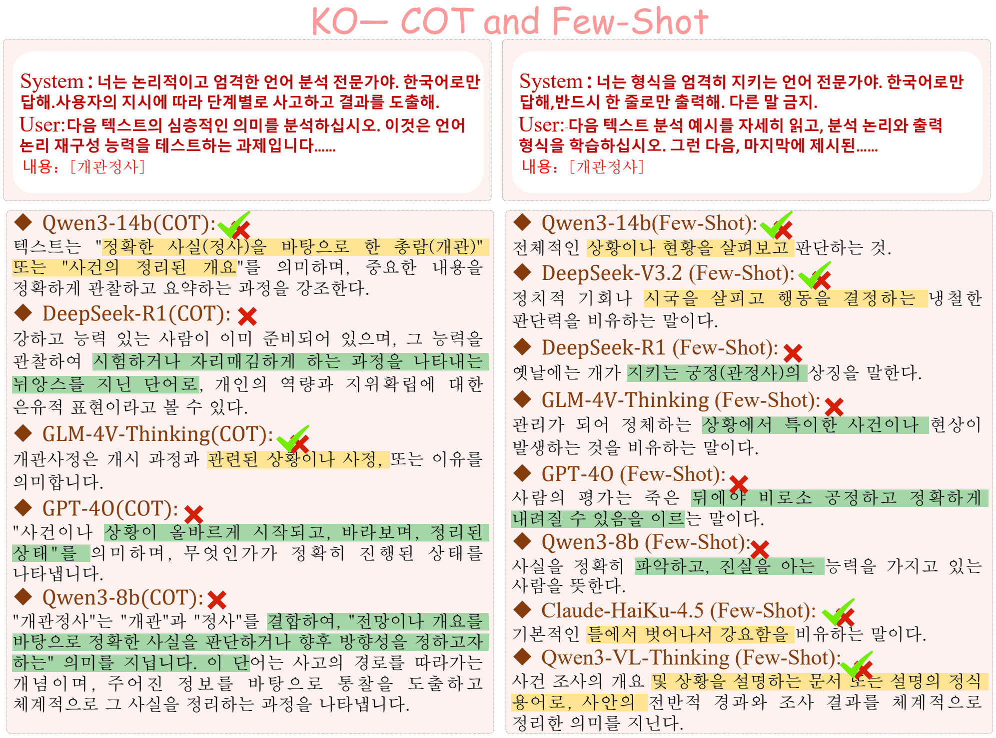
◆ **DeepSeek-R1 (Few-Shot):** ✓✗
蛇か龍を描き終えた後、その中の目をまるでかすかに瞬くように点入れて完成させることに、比喩的に熱心に努力する姿を表す言葉である。

◆ **Gemini-2.5-flash (Few-Shot):** ✓
竜を描き終えて最後に瞳を点入れると、竜が生きて動き出したという故事に由来する。ある事柄の最も重要な部分を完成させるか、あるいは文章や言葉の要点をつきつけることで全体を生き生きとさせることを比喩する。

◆ **GLM-4V-Thinking (Few-Shot):** ✓
ドラゴンの描き上げを終えた後、続けて目を点で表すという行為から転じて……

◆ **GPT-4O (Few-Shot):** ✓✗
既にうまくやった事柄を締めくくることで、さらに完成度を高めることを比喩する言葉である。

Figure 12: **Comparison of prompting strategies on Japanese.** This detailed analysis highlights the syntax recovery challenges in mixed script inputs.



KO— COT and Few-Shot

System：너는 논리적이고 엄격한 언어 분석 전문가야. 한국어로만 답해.사용자의 지시에 따라 단계별로 사고하고 결과를 도출해.
User:다음 텍스트의 심층적인 의미를 분석하십시오. 이것은 언어 논리 재구성 능력을 테스트하는 과제입니다......
내용：［개관정사］

◆ **Qwen3-14b(COT):** ✓✗
텍스트는 "정확한 사실(정사)을 바탕으로 한 총람(개관)" 또는 "사건의 정리된 개요"를 의미하며, 중요한 내용을 정확하게 관찰하고 요약하는 과정을 강조한다.

◆ **DeepSeek-R1(COT):** ✗
강하고 능력 있는 사람이 이미 준비되어 있으며, 그 능력을 관찰하여 시험하거나 자리매김하게 하는 과정을 나타내는 뉘앙스를 지닌 단어로, 개인의 역량과 지위확립에 대한 은유적 표현이라고 볼 수 있다.

◆ **GLM-4V-Thinking(COT):** ✓✗
개관사정은 개시 과정과 관련된 상황이나 사정, 또는 이유를 의미합니다.

◆ **GPT-4O(COT):** ✗
"사건이나 상황이 올바르게 시작되고, 바라보며, 정리된 상태"를 의미하며, 무엇인가가 정확히 진행된 상태를 나타냅니다.

◆ **Qwen3-8b(COT):** ✗
"개관정사"는 "개관"과 "정사"를 결합하여, "전망이나 개요를 바탕으로 정확한 사실을 판단하거나 향후 방향성을 정하고자 하는" 의미를 지닙니다. 이 단어는 사고의 경로를 따라가는 개념이며, 주어진 정보를 바탕으로 통찰을 도출하고 체계적으로 그 사실을 정리하는 과정을 나타냅니다.

System：너는 형식을 엄격히 지키는 언어 전문가야. 한국어로만 답해,반드시 한 줄로만 출력해. 다른 말 금지.
User:다음 텍스트 분석 예시를 자세히 읽고, 분석 논리와 출력 형식을 학습하십시오. 그런 다음, 마지막에 제시된......
내용：［개관정사］

◆ **Qwen3-14b(Few-Shot):** ✓✗
전체적인 상황이나 현황을 살펴보고 판단하는 것.

◆ **DeepSeek-V3.2 (Few-Shot):** ✓✗
정치적 기회나 시국을 살피고 행동을 결정하는 냉철한 판단력을 비유하는 말이다.

◆ **DeepSeek-R1 (Few-Shot):** ✗
옛날에는 개가 지키는 궁정(관정사)의 상징을 말한다.

◆ **GLM-4V-Thinking (Few-Shot):** ✗
관리가 되어 정체하는 상황에서 특이한 사건이나 현상이 발생하는 것을 비유하는 말이다.

◆ **GPT-4O (Few-Shot):** ✗
사람의 평가는 죽은 뒤에야 비로소 공정하고 정확하게 내려질 수 있음을 이르는 말이다.

◆ **Qwen3-8b (Few-Shot):** ✗
사실을 정확히 파악하고, 진실을 아는 능력을 가지고 있는 사람을 뜻한다.

◆ **Claude-HaiKu-4.5 (Few-Shot):** ✓✗
기본적인 틀에서 벗어나서 강요함을 비유하는 말이다.

◆ **Qwen3-VL-Thinking (Few-Shot):** ✓✗
사건 조사의 개요 및 상황을 설명하는 문서 또는 설명의 정식 용어로, 사안의 전반적 경과와 조사 결과를 체계적으로 정리한 의미를 지닌다.

Figure 13: **Comparison of prompting strategies on Korean.** We observe distinct behaviors between standard prompting and reasoning enhanced generation.

Language	Type	Baseline Prompt Template (Visualized Layout)
ZH-CN	System	你是一个严格遵循格式要求的语言专家。请你回答为中文。请只输出一行，不要多余空行或其他文字： 判断含义：**[解释]**
	User	请你判断这个的含义。直接判断并说出含义，不要有其他废话，尽可能全面： 判断含义：**[解释]** 成语：{idiom}
ZH-TW	System	你是一個嚴格遵循格式要求的語言專家。請用繁體中文回答。請只輸出一行，不要多餘空行或其他文字： 判斷含義：**[解釋]**
	User	請你判斷這個的含義。直接判斷並說出含義，不要有其他贅詞，可能全面： 判斷含義：**[解釋]** 成語：{idiom}
KO	System	너는 형식을 엄격히 지키는 언어 전문가다. 한국어로만 답하라. 반드시 한 줄로만 출력하고 다른 문장은 포함하지 마라: **의미: [설명]**
	User	다음에 제시된 표현의 의미를 한 줄로 설명하라. 군말은 포함하지 말고 가능한 한 완전하게 설명하라: **의미: [설명]** 내용:{idiom}
JA	System	あなたは形式を厳守する言語専門家です。日本語で回答してください。必ず一行のみを出力し余分な文言や空行は含めないでください： 意味：**[説明]**
	User	以下に示す表現の意味を一行で説明してください。余分な文言は禁止し可能な限り完全に述べてください： 意味：**[説明]** 内容：{idiom}

Table 8: Visualized baseline prompt templates used for idiom meaning inference across four languages. Prompts are shown in their formatted layout to illustrate the structural constraints imposed on model outputs.

Language	Type	Few Shot Prompt Template Visualized Layout
ZH-CN	System	你是一个严谨的逻辑分析专家。请严格遵循用户的指令格式输出结果。禁止输出任何开场白或结束语。
	System	start deep thinking, please always use the thinking mode
	User	请仔细阅读以下文本分析示例，学习其分析逻辑和输出格式。然后，对最后给出的待分析文本进行同样的分析。 示例1 输入文本：惊人一鸣 Thinking Process：分析字面，惊人一鸣在语义上不通顺。观察字词组合，发现一鸣通常作为主语或状语，惊人作为结果。调整语序为一鸣惊人，符合一声鸣叫让人震惊的逻辑。 Final Meaning：比喻平时没有突出的表现，一下子做出惊人的成绩。 示例2 输入文本：添足画蛇 Thinking Process：分析字面，添足意味着增加脚，画蛇是动作主体。逻辑上先有画蛇，后有添足的多余行为。重组为画蛇添足。 Final Meaning：比喻做了多余的事，非但无益，反而不合适。 示例3 输入文本：求剑刻舟 Thinking Process：分析字面，求剑是目的，刻舟是手段。按照因果和动作发生的时间顺序，应先在船上刻记号，再去寻找剑。重组为刻舟求剑。 Final Meaning：比喻拘泥不知变通，不懂得根据实际情况处理问题。 待分析文本 输入文本：{idiom} 请严格模仿上述示例的逻辑和格式输出，必须严格两行。
ZH-TW	System	你是一個嚴謹的邏輯分析專家。請嚴格遵循用戶的指令格式輸出結果。禁止輸出任何開場白或結束語。
	System	start deep thinking, please always use the thinking mode
	User	請仔細閱讀以下文本分析範例，學習其分析邏輯和輸出格式。然後，對最後給出的待分析文本進行同樣的分析。 範例1 輸入文本：驚人一鳴 Thinking Process：分析字面，驚人一鳴在語義上不通順。觀察字詞組合，發現一鳴通常作為主語或狀語，驚人作為結果。調整語序為一鳴驚人，符合一聲鳴叫讓人震驚的邏輯。 Final Meaning：比喻平時沒有突出的表現，一下子做出驚人的成績。 範例2 輸入文本：添足畫蛇 Thinking Process：分析字面，添足意味著增加腳，畫蛇是動作主體。邏輯上先有畫蛇，後有添足的多餘行為。重組為畫蛇添足。 Final Meaning：比喻做了多餘的事，非但無益，反而不合適。 範例3 輸入文本：求劍刻舟 Thinking Process：分析字面，求劍是目的，刻舟是手段。按照因果和動作發生的時間順序，應先在船上刻記號，再去尋找劍。重組為刻舟求劍。 Final Meaning：比喻拘泥不知變通，不懂得根據實際情況處理問題。 待分析文本 輸入文本：{idiom} 請嚴格模仿上述範例的邏輯和格式輸出，必須嚴格兩行。
KO	System	너는 논리적이고 엄격한 언어 분석 전문가야. 한국어로만 답해. 사용자의 지시에 따라 단계별로 사고하고 결과를 도출해.
	System	start deep thinking, please always use the thinking mode
	User	다음 텍스트 분석 예시를 자세히 읽고 분석 논리와 출력 형식을 학습하십시오. 그런 다음 마지막에 제시된 분석 대상 텍스트에 대해 동일한 분석을 수행하십시오. 예시 1 입력 텍스트: 경인일명 (驚人一鳴) Thinking Process: 새가 한 번 울어 사람들을 놀라게 한다는 인과 관계가 자연스럽다. 따라서 어순을 일명경인으로 재구성한다. Final Meaning: 평소에는 조용하다가 갑자기 놀라운 성과를 내어 세상을 놀라게 함을 비유하는 말이다. 예시 2 입력 텍스트: 첨족화사 (添足蛇) Thinking Process: 뱀을 그리고 나서 발을 더하는 행위가 불필요하다. 어순을 화사첨족으로 재구성한다. Final Meaning: 쓸데없는 일을 덧붙여서 도리어 일을 그르침을 이르는 말이다. 예시 3 입력 텍스트: 구검각주 (求劍刻舟) Thinking Process: 칼을 찾기 위한 수단으로 배에 표시를 하므로 먼저 각주하고 나중에 구검한다. 어순을 각주구검으로 재구성한다. Final Meaning: 변화하는 상황을 모르고 낡은 방식에 집착하는 어리석음을 비유한다. 분석 대상 텍스트 입력 텍스트: {idiom} 위 예시들의 논리와 형식을 엄격히 모방하여 출력하십시오. 반드시 두 줄만 출력하십시오.
JA	System	あなたは厳格な論理を持つ言語分析の専門家です。ユーザーの指示するフォーマットに厳密に従い、余計な挨拶や結びの言葉は出力しないでください。
	System	start deep thinking, please always use the thinking mode
	User	次のテキスト分析例を注意深く読み、分析の論理と出力形式を学習してください。その後、最後に示される分析対象テキストについて同様の分析を行ってください。 例 1 入力テキスト：驚人一鳴 Thinking Process: 驚人一鳴は不自然であり、一鳴が原因で驚人が結果となるため語順を一鳴驚人に調整する。 Final Meaning: 普段は目立たないが突然すばらしい成果を出して人々を驚かせることのたとえ。 例 2 入力テキスト：添足画蛇 Thinking Process: まず蛇を描き、その後で足を足す行為が余計である。語順を画蛇添足に再構成する。 Final Meaning: 余計なことをしてかえって不適切になることのたとえ。 例 3 入力テキスト：求剣刻舟 Thinking Process: 剣を探すために舟に印を刻すので先に刻舟し後に求剣する。語順を刻舟求剣に再構成する。 Final Meaning: 状況の変化を理解せず古い方法に固執する愚かさのたとえ。 分析対象テキスト 入力テキスト：{idiom} 上記の例の論理と形式を厳密に模倣して出力してください。必ず二行のみを出力してください。

Table 9: Visualized few shot prompt templates used for multilingual idiom analysis. Prompts are shown in formatted layout to illustrate in context reasoning patterns and the strict two line output constraint.

Language	Type	CoT Prompt Template Visualized Layout
ZH-CN	System	你是一个严谨的逻辑分析专家。请严格遵循用户的指令格式输出结果。禁止输出任何开场白或结束语。 start deep thinking, please always use the thinking mode
	User	请分析以下文本的深层含义。这是一个测试语言逻辑重组能力的任务。 为了确保准确性，请务必一步步思考，严格遵循以下分析路径。 1 字义拆解分析文本中每个字的独立含义。 2 结构重组观察这些字之间可能存在的句法逻辑，判断是否存在语序错位。 3 语义推导基于重组后的逻辑，推导该文本的整体寓意。 请严格按照以下特定格式输出，必须严格两行。 Thinking Process: 在这里写下你的详细分步推理过程，包括对字序和结构的分析。 Final Meaning: 基于推理得出的最终含义解释，只给结论，不要再解释推理过程，尽可能全面。 输入文本：{idiom}
ZH-TW	System	你是一個嚴謹的邏輯分析專家。請嚴格遵循用戶的指令格式輸出結果。禁止輸出任何開場白或結束語。 start deep thinking, please always use the thinking mode
	User	請分析以下文本的深層含義。這是一個測試語言邏輯重組能力的任務。 為了確保準確性，請務必一步步思考，嚴格遵循以下分析路徑。 1 字義拆解分析文本中每個字的獨立含義。 2 結構重組觀察這些字之間可能存在的句法邏輯，判斷是否存在語序錯位。 3 語義推導基於重組後的邏輯，推導該文本的實際整體寓意。 請嚴格按照以下特定格式輸出，必須嚴格兩行。 Thinking Process: 在這裡寫下你的詳細分步推理過程，包括對字序和結構的分析。 Final Meaning: 基於推理得出的最終含義解釋，只給結論，不要再解釋推理過程，盡可能全面。 輸入文本：{idiom}
KO	System	너는 논리적이고 엄격한 언어 분석 전문가야. 한국어로만 답해. 사용자의 지시에 따라 단계별로 사고하고 결과를 도출해. start deep thinking, please always use the thinking mode
	User	다음 텍스트의 심층적인 의미를 분석하십시오. 이것은 언어 논리 재구성 능력을 테스트하는 과제입니다. 정확성을 보장하기 위해 반드시 단계별로 사고하고 다음 분석 경로를 엄격히 따르십시오. 1 글자 풀이 텍스트에 포함된 각 글자의 개별적인 의미를 분석합니다. 2 구조 재구성 글자들 간의 잠재적인 구문 논리를 관찰하여 어순이 뒤섞여 있는지 판단합니다. 3 의미 도출 재구성된 논리를 바탕으로 텍스트의 전체 의미를 추론합니다. 다음 형식에 맞춰 엄격히 출력하십시오. 반드시 두 줄이어야 하며 줄을 추가하거나 빼지 마십시오. Thinking Process: 여기에 글자 순서와 구조 분석을 포함한 상세한 단계별 추론 과정을 작성하십시오. Final Meaning: 추론을 통해 도출된 최종 의미 설명. 결론만 제시하고 추론 과정을 반복하지 마십시오. 최대한 포괄적으로 작성하십시오. 입력 텍스트: {idiom}
JA	System	あなたは厳格な論理を持つ言語分析の専門家です。ユーザーの指示するフォーマットに厳密に従い、余計な挨拶や結びの言葉は出力しないでください。 start deep thinking, please always use the thinking mode
	User	以下のテキストの深層的な意味を分析してください。これは言語論理の再構築能力をテストするタスクです。 正確性を確保するため必ずステップバイステップで思考し以下の分析パスを厳守してください。 1 文字の解読 テキストに含まれる各文字の個別の意味を分析します。 2 構造の再構築 文字間の構文論理を観察し語順が入れ替わっていないか判断します。 3 意味の導出 再構築された論理に基づいてテキスト全体の意味を推論します。 以下の形式に厳密に従って出力してください。必ず二行でなければなりません。 Thinking Process: ここに文字の順序と構造の分析を含む詳細な段階的推論プロセスを記述してください。 Final Meaning: 推論に基づいて導き出された最終的な意味の説明。結論のみを提示し推論過程を繰り返さないでください。可能な限り包括的に記述してください。 入力テキスト：{idiom}

Table 10: Visualized CoT prompt templates used for multilingual idiom analysis. The prompts enforce a fixed two line output format and instruct stepwise reasoning for semantic and structural reconstruction.

Model Family	S_{Acc}^{mean}	S_{Cons}	S_{Logic}	S_{Rob}	S_{Info}	**Recovery (%)**
Tier 1: High-Performance Reasoning Models						
Qwen-3-14B	0.862	0.857	0.873	0.904	0.841	**100**
Qwen-3-VL$_{\text{Think}}$	0.884	0.853	0.886	0.913	0.832	**100**
Gemini-2.5-Flash	0.871	0.864	0.859	0.898	0.821	**100**
Tier 2: Mid-Range General Models						
GPT-4o	0.879	0.882	0.858	0.914	0.843	**100**
GLM-4V$_{\text{Think}}$	0.833	0.842	0.845	0.883	0.801	**100**
DeepSeek-V3.2	0.874	0.868	0.893	0.909	0.847	**100**
Tier 3: Efficient & Baseline Models						
DeepSeek-R1	0.821	0.848	0.824	0.879	0.793	**100**
Claude-HaiKu-4.5	0.843	0.861	0.834	0.887	0.802	**100**
Gemma-3-4B	0.783	0.819	0.802	0.847	0.742	**100**
Llama-3-8B	0.758	0.832	0.791	0.853	0.723	**100**
Llama-2-7B	0.742	0.814	0.774	0.838	0.703	**100**
Qwen-3-8B	0.779	0.823	0.788	0.852	0.739	**100**

Table 11: **Reference Diagnostic Metrics under Unperturbed Inputs.** This appendix table reports diagnostic component scores obtained from a reference evaluation run using unperturbed inputs and identical inference configurations as the main experiments. Since no perturbation is applied, the recovery rate is uniformly set to 100% for all models. These results are provided for contextual reference only and are not used in any comparative or statistical claims.